МІНІСТЕРСТВО ОСВІТИ і НАУКИ УКРАЇНИ
Київський національний університет будівництва і архітектури

СТАТИСТИКА В УПРАВЛІННІ ЕКОНОМІКОЮ БУДІВНИЦТВА І НЕРУХОМОСТІ

Рекомендовано вченою радою Київського національного університету будівництва і архітектури як навчальний посібник для студентів галузі знань 05 «Соціальні та поведінкові науки» спеціальності 051 «Економіка»

Київ 2022

УДК 69.003.1 (083)
С45

Автори: Л.В. Сорокіна, д-р екон. наук, професор (передмова, вступ, п. 1.3.2, 1.3.3, підрозділи 1.4, 2.1, 2.2, 2.3, 3.1, 3.2, 3.4, завдання для самостійного опрацювання матеріалу; література; загальне керівництво роботою);
А.Ф. Гойко, канд. екон. наук, професор (п. 1.1.4, 1.3.1);
С.П. Стеценко, д-р екон. наук, професор (п. 1.1.1, 1.1.2, 1.1.3);
І.О. Шапошникова, канд. екон. наук, доцент (п. 3.3.1, 3.3.2);
В.Я. Гаврилюк, канд. екон. наук, доцент (п. 1.2.1, 1.2.2)

Рецензенти: *О.К. Щетініна*, д-р фіз.-мат. наук, професор, КНТЕУ;
К.В. Павлов, д-р екон. наук, професор, ВНУ ім. Лесі Українки;
О.Ю. Бєлєнкова, д-р екон. наук, професор, КНУБА

Затверджено на засіданні вченої ради Київського національного університету будівництва і архітектури, протокол № 45 від 29 жовтня 2021 року.

Статистика в управлінні економікою будівництва і нерухомості: навч. посіб. / Л.В. Сорокіна та ін. – Київ: КНУБА, 2021. – 168 с.

С45

ISBN 978-966-627-239-6

Розглянуто проблеми застосування методів статистичного аналізу даних в процесі розв’язування наукових та практичних задач, що виникають під час управління розвитком будівельних підприємств та дослідження ринків нерухомості. Наведено приклади використання статистичних методів для розв’язання практичних ситуацій. Містить задачі і вправи для самостійного опрацювання.

Призначено для студентів галузі знань 05 «Соціальні та поведінкові науки» спеціальності 051 «Економіка», денной та заочної форм навчання.

УДК 69.003.1 (083)

ISBN 978-966-627-239-6

ЗМІСТ

Присвячується пам'яті
Юрія Васильовича Четверікова,
кандидата економічних наук, доцента,
доброї, чуйної людини,
талановитого викладача
і натхненника цієї книги

ПЕРЕДМОВА

Ініціатором й натхненником представленої праці був доцент кафедри економіки будівництва КНУБА, кандидат економічних наук Юрій Васильович Четверіков, але, на превеликий сум і жаль, він так і не встиг побачити надрукований примірник книги. І тому посібник присвячено його світлій пам'яті.

Необхідність написання навчального посібника зумовлена потребою й у теоретичних знаннях, й у методиці застосування набутих знань на практиці. У пропонованому виданні узагальнено більш ніж двадцятирічний досвід викладання статистики у вищій школі. Воно сприятиме засвоєнню не лише знань зі статистики, а й розумінню методів виконання статистичного аналізу реальних явищ та процесів, пов'язаних з економікою будівництва та ринком нерухомості, зокрема з питаннями економічного управління, вартісного інжинірингу, проведення маркетингових чи наукових досліджень.

Колектив авторів висловлює щиру вдячність рецензентам та керівництву університету за цінні зауваження та всіляке сприяння цьому виданню.

Автори з повагою та вдячністю розглянуть усі зауваження, пропозиції та побажання читачів і обов'язково врахують їх у подальших виданнях.

Число являє
усю сутність кожної речі“
Платон «Теєтет»

ВСТУП

Вагомою конкурентної перевагою сучасного фахівця-економіста є вміння аналізувати великі масиви інформації, виробляючи на їхній основі найефективніші рішення. Насамперед цінується здатність результативної роботи із кількісними даними, які є головним чинником формування стратегічних планів, розроблення тактичних заходів обраної економічної політики, контролю виконання операційних бюджетів. Визначальна роль статистичної підготовки майбутніх економістів та інженерів посилюється в умовах тотальної примусової цифровізації усіх господарських процесів, оскільки цінний для підприємства працівник має бути не лише оператором, що вводить первинні дані до комп'ютера. З огляду на те, що економічне управління підприємством — особлива галузь застосування набутих теоретичних знань, зацікавленим у довгостроковому зростанні бізнесу стейкхолдерам-працівникам доводиться постійно розробляти нові та вдосконалювати існуючі он-лайн застосунків та форми. Адже наразі лише з їхньою допомогою стає можливим глибоко вивчити смаки та уподобання клієнтів й сформувати способи їхнього задоволення. В цьому зв'язку корисними будуть знання зі статистики та вміння виконувати статистичні розрахунки.

Діджиталізація не оминула й підприємства будівництва, оскільки необхідність систематичного моніторингу цін ресурсів, вивчення ефективності новітніх будівельних технологій та способів організації процесів створення й управління нерухомістю, досліджень міцності, надійності, довговічності будівельних матеріалів, виробів конструкцій та економічного ефекту від їх використання потребують також вмінь здійснювати кількісний аналіз емпіричної інформації, узагальнюючи виявлені залежності у вигляді правил чи формул. Окремий сегмент економічної діяльності, ефективність якої визначається уміннями працювати з великими масивами числових даних, являє собою управління нерухомістю та оцінювання її вартості. Тому глибокі якісні знання зі статистики є необхідними і для вартісних інженерів, інженерів-консультантів, кошторисників, ріелторів, оцінювачів.

Для закріплення набутих теоретичних знань після кожного розділу вміщено комплекс практичних завдань, виконання яких сприяє формуванню стійких вмінь та навичок кількісного аналізу великих масивів економічної інформації, джерелами якої є не лише дані спостережень господарських процесів будівельних підприємств, але й публічна інформація щодо розвитку національної економіки та світових господарських процесів.

Мета дисципліни – дати студентам знання основ статистичного вимірювання, методів узагальнення та аналізу інформації про соціально-економічні явища і процеси, про закономірності суспільного життя. Студент повинен знати методи розрахунку найважливіших статистичних показників, розуміти їх сутність, вміти збирати, обробляти та аналізувати інформацію, виявляти й оцінювати закономірності формування, розвитку та взаємодії складних за своєю природою соціально-економічних явищ і процесів.

Завдання дисципліни полягає у виробленні у студентів таких умінь та навичок:

- оволодіння методичним інструментарієм статистичного аналізу суспільно-економічних явищ та процесів;
- оволодіння методичним інструментарієм вивчення тенденцій розвитку економічних суб'єктів;
- формування вмінь використання програмного забезпечення (MS Excel, Matlab, Statistica) під час статистичного аналізу даних;
- формування практичних навичок для створення програмних продуктів, спрямованих на автоматизацію розрахункових операцій.

У результаті вивчення дисципліни студент повинен **знати:**

- методологію побудови статистичних групувань;
- методи аналізу рядів розподілу;
- правила визначення обсягу вибірки та помилок вибірки;
- методи перевірки статистичних гіпотез;
- статистичні методи вивчення взаємозв'язків явищ.

У результаті вивчення дисципліни студент повинен **вміти:**

- застосовувати програмні засоби комп'ютерної техніки для розрахунку статистичних показників;
- проводити статистичні дослідження та експерименти із встановленням значущості отриманих результатів;
- застосовувати статистичні методи вимірювання взаємозв'язків;
- будувати прогнозні моделі розвитку економічних суб'єктів на основі виявлення та вимірювання тенденцій розвитку та сезонних коливань.

Розділ 1
СТАТИСТИЧНІ ПОКАЗНИКИ І АНАЛІЗ РЯДІВ РОЗПОДІЛУ В ДОСЛІДЖЕННЯХ ЕКОНОМІКИ БУДІВНИЦТВА

1.1. СТАТИСТИЧНІ ПОКАЗНИКИ, ВИДИ, МЕТОДИКА РОЗРАХУНКУ

1.1.1. Основні теоретичні відомості

Предметом статистики є вивчення кількісного співвідношення масових явищ та процесів у нерозривному зв'язку з їх якісним змістом за певних умов місця та часу.

Статистична закономірність – *закономірність, в якій необхідність пов'язана в кожному окремому явищі з випадковістю і лише в сукупності явищ виявляє себе як закон.*

Групи методів статистики:

- методи масового спостереження;
- методи зведення та групування;
- методи визначення узагальнювальних і синтетичних показників (методи середніх та відносних величин, аналіз рядів розподілу, вимірювання зв'язку).

Статистичний показник – число в сукупності з набором ознак, що характеризують обставини, яких воно стосується (що, де, коли, яким чином підлягає *вимірюванню).*

Статистичні дані – сукупність показників, отриманих як результат статистичного спостереження або оброблення даних.

Статистична сукупність – множина елементів, поєднаних спільними умовами та причинами. Сукупність складається з окремих одиниць, які мають спільні риси або ознаки. Коливання значень ознаки в сукупності називають варіацією.

Етапи статистичного дослідження:

- збирання статистичних даних та їхнє первісне оброблення (чищення);
- зведення, групування, обчислення середніх і відносних величин;
- аналіз варіації, взаємозв'язку, перевірка гіпотез.

Статистичний показник – число в сукупності з набором ознак, що характеризують обставини, яких воно стосується (що, де, коли, яким чином підлягає вимірюванню). На рис. 1.1. представлено п'ять способів класифікації статистичних показників.

Рис. 1.1. Критерії та способи класифікації статистичних показників

Для практичних розрахунків найбільш корисною є класифікація **за формою вираження**. Це останній блок на рис. 1.1, у ньому виділено три види статистичних показників:

- абсолютні величини;
- відносні величини;
- середні величини.

Розглянемо докладніше способи й особливості визначення кожного з наведених видів показників.

1.1.2. Поняття про абсолютні величини та їх види

Абсолютні величини відображають розміри, рівні, обсяги масових соціально-економічних явищ у конкретних умовах місця та часу. **За ступенем охоплення** виокремлюють такі види абсолютних величин:

- **індивідуальні,** які характеризують окремі одиниці сукупності **(**зарплата одного робітника в гривнях, витрати на комунальні послуги одного домогосподарства);
- **групові** – відображають розміри одиниць сукупності (протяжність автодоріг державного значення);
- **сумарні** – характеризують обсяг статистичної сукупності.

Абсолютні статистичні величини завжди є **іменованими.** Залежно від сутності досліджуваного явища і мети аналізу виокремлюють такі види абсолютних величин:

- **натуральні** одиниці вимірювання відображають певні притаманні явищам природні і споживчі властивості і виражаються у відповідних одиницях ваги, довжини, площі, об'єму. Випуск напоїв — у декалітрах, обсяг зведеного житла – у квадратних метрах, будівництво доріг — у кілометрах, виготовлення сантехнічних виробів – у штуках. Натуральні показники можуть бути:
- **прості**;
- **комбіновані**, які являють собою поєднання декількох різнойменних одиниць вимірювання, застосовуваних для всебічного визначення розмірів деяких явищ. Наприклад, робота транспорту: тонно-кілометри, пасажиро-кілометри, крісло-кілометри; баланс робочого часу: людино-години, машино-години; виробництво і споживання електроенергії: кіловат-години. Комбіновані натуральні одиниці відображають відразу декілька сторін досліджуваного явища;
- **умовно-натуральні** одиниці вимірювання застосовують для визначення загального обсягу виробництва однорідних продуктів, які відрізняються своїми споживчими властивостями, тому підсумовуванням їхнього обсягу в натуральних одиницях вимірювання було б некоректним з погляду статистичного аналізу. Загальний обсяг виробництва в умовно-натуральному вимірюванні (V) визначають за формулою:

$$V = \sum_{i=1}^{n} Q_i k_i \text{ ,} \qquad (1.1)$$

де Q_i – фізичний обсяг окремих різновидів продукції;

k_i – коефіцієнт перерахунку, який обчислюють за співвідношенням різновидів такого явища, на один різновид, взяти за еталон.

Приклад 1.1*. Завод будівельних матеріалів виготовив 1 млн шт. цегли марки М100, 600 тис. шт. цегли марки М150, 300 тис. шт. цегли марки М200. Визначити загальний обсяг випуску цегли у перерахунку на умовну цеглу (якою є цегла М100). Коефіцієнти перерахунку становлять:* k_1*=1=100/100,* $k_{1,5}$*=1,5=150/100,* k_2*=2=200/100:*

$$V = \sum_{i=1}^{n} Q_i k_i = 1 \cdot 1 + 0{,}6 \cdot 1{,}5 + 0{,}3 \cdot 2 = 2{,}5 \text{ млн шт. умовної цегли}$$

У вітчизняній статистиці застосовують такі умовні натуральні одиниці вимірювання:

- умовне паливо, теплоємність якого становить 29,3076 МДж, наприклад:
- 100 т торфу еквівалентні 81,9 т умовного палива;
- 100 т нафти еквівалентні 153,6 т умовного палива;

- умовні консервні банки, об’ємом 353,4 $см^3$ (№8);
- умовне поголів’я (за умовну голову взято особину дорослої рогатої худоби) застосовують для загальної характеристики стану і розвитку тваринництва;
- кормова одиниця (к.о.) застосовується для узагальнювального обліку обсягу різних кормів залежно від їхньої поживної цінності;
- умовний вагон, що за своїми властивостями є відповідним 4-вісному залізничному вагону;
- умовний вміст поживних речовин у мінеральних добривах різних видів (у перерахунку на 100% поживних речовин);
- умовний вміст активної речовини у хімічних засобах захисту рослин різних видів (у 100%-му обчисленні за активною речовиною);
- вартісні (грошові) одиниці вимірювання дають змогу визначати розміри різних за своїм змістом явищ, не порівнюваних у натуральних одиницях вимірювання. Такими одиницями може бути національна або іноземна валюта, однак недоліком цих одиниць вимірювання є мінливість цін благ, що призводить до переоцінювання вартісних показників у ціни одного й того самого періоду, що називають порівнянними цінами (інфлювання, дефілювання): наприклад, порівнянні ціни 2011 року;
- трудові одиниці вимірювання, потрібні для визначення обсягу затрат робочого часу й оцінювання ефективності його використання можуть бути простими або комбінованими. Наприклад, затрати робочого часу на виробництво продукції можуть бути виражені в чисельності робітників, залучених до виробництва (проста натуральна одиниця вимірювання), людино-днях, людино-годинах (комбіновані натуральні одиниці вимірювання).

Для планування капітальних вкладень у розвиток виробничої потужності умовно-натуральні показники відіграють визначну роль. Для прикладу розглянемо особливості використання такого умовно-натурального показника, як умовна банка.

Консервовані продукти випускають в бляшаній, скляній, полімерній. дерев'яній і картонній тарі, яка, крім того, відрізняється формою, розмірами. місткістю. Зважаючи на різноманітність тари, застосовуваної для консервування продуктів, а також для зручності планування, обліку, звітності розроблено систему обчислення консервованої продукції в умовних одиницях.

Одиницями обчислення консервної продукції є облікові, або умовні, банки, а також масові одиниці – кілограми або тони (для солоної, квашеної, або замороженої продукції, сушених фруктів, овочів і різноманітних напівфабрикатів).

Для обчислення готової продукції в облікових одиницях застосовують два види умовних банок: об'ємні і масові.

Умовна ОБ'ЄМНА банка – це бляшанка № 8 місткістю 353,4 мл.

Умовна МАСОВА банка містить 400 г продукту.

В об'ємних умовних банках обліковують усі консервовані фрукти, овочі, м'ясо, рибу, молоко.

У масових умовних банках обліковують варення, джеми, повидло, желе, маринади, фруктові та овочеві соки, соуси, пюре.

Для визначення кількості умовних банок в тій чи іншій тарі треба повний об'єм цієї тари розділити на 353,4 мл. Для визначення масових умовних банок слід масу продукту розділити на 400 г.

Обліковуючи умовні банки для консервованої продукції, обов'язково виконують перерахунок на 12% сухих речовин. У випадку концентрованого томатного соку перераховують на 5% сухих речовин.

Перерахунок на умовні банки здійснюють за формулою

$$n_{уб} = G_n \frac{C_n}{C_{уб}} 0{,}4, \qquad (1.2)$$

де $n_{уб}$ – кількість умовних банок, шт.; $G_п$ – маса концентрованого соку, кг; $C_п$ – концентрація концентрованого продукту; $C_{уб}$ – концентрація продуктів в умовних банках, % (12% або 5%).

Концентровані фруктові продукти (пасти, соуси, соки) переводять в умовні банки, помноживши на масову одиницю (0,4 кг) концентрованого продукту на коефіцієнт, який залежить від вмісту сухих речовин (табл. 1.1).

Таблиця 1.1

Коефіцієнт переведення масової одиниці в умовні банки

Найменування продукту	Вміст сухих речовин, %	Коефіцієнт
Фруктовий соус	32	1,5
Фруктова паста	18	1,5
	25	2,0
	30	2,5
Сік мандариновий	45	4,5
Сік яблучний	55	5,0

Для зручності і швидкого перерахунку фізичних банок в умовні і навпаки для кожного виду бляшаної і скляної тари визначено перевідні об'ємні коефіцієнти, їх можна знайти у довідковій літературі із харчових технологій. Для переведення фізичних банок в умовні потрібно їх кількість помножити на відповідний коефіцієнт, для переведення умовних банок у фізичні – кількість умовних банок поділити на коефіцієнт.

Для переведення в умовні банки продукції, яку обліковують за масовим перевідним коефіцієнтом, потрібно знати масу нетто продукту в кожному

фасуванні, а для концентрованих томат-продуктів, соків. фруктового пюре, крім того, – ще й фактичний вміст сухих речовин у готовому продукті.

У технологічних інструкціях для багатьох консервів рецептура і норми витрат сировини наведено на 1 т готового продукту. Для перерахунку 1 т консервів в умовні банки застосовують різні методи, незалежно від того, яку умовну банку (масову чи об'ємну) взято для певного виду консервів. Для таких консервів, як томат-продукти, фруктові і овочеві соки, маринади, повидло, джеми та інші, за умовну банку беруть масу 400 г. Перерахунок виконують, ділячи 1000 кг консервів на 0,4 кг. У разі концентрованої продукції враховують вміст сухих речовин в умовній банці – 12%, а для концентрованого соку – 5%.

Один із способів перерахунку 1 т консервів в об'ємні умовні банки полягає в тому, що масу нетто продукту фізичної банки ділять на встановлений для неї перевідний коефіцієнт, відтак 1000 кг ділять на одержану масу умовної банки.

Норму витрат сировини і матеріалів на 1 т консервів перераховують на одну тисячу умовних банок за формулами:

$$H_{уб} = G_{тк} \frac{M}{1000} K; \tag{1.3}$$

$$H_{уб} = G_{тк} \bullet M', \tag{1.4}$$

де $H_{уб}$ – норма витрат на 1 тис. умовних банок, кг; $G_{тк}$ – норма витрат на 1 т консервів, кг; M – маса нетто тисячі фізичних банок, кг; K – перевідний коефіцієнт для фізичної банки, кг; M' – маса нетто однієї умовної банки, кг.

У випадках, коли норми витрат сировини і матеріалів визначено для 1 т готового продукту, а їх потрібно перерахувати на один туб консервів, для яких маса умовної банки дорівнює 400 г, треба масу продукту за рецептурою поділити на 1 000 і помножити на 400. Але простіше відразу масу продукції за рецептурою ділити на 2,5 (2,5=1000 / 400).

Наведені формули (1.2) – (1.4) потрібно використовувати, складаючи технічне завдання для проєктування підприємств переробної промисловості.

1.1.3. Поняття про відносні величини і форми їх вираження

Відносні величини – це показники, які характеризують кількісні співвідношення, притаманні суспільним явищам і процесам, що є відношенням двох статистичних величин. *Кожна відносна величина є дробом, де*

чисельник – ***порівнювана величина, а знаменник*** – ***база порівняння (основа відносної величини****, тобто величина, з якою порівнюють, взята дослідником за еталон). База порівняння може дорівнювати 1, 100, 1000 тощо.* Залежно від того, яке числове значення має база порівняння, розрізняють:

- **індекси;**
- **коефіцієнти** – застосовують, коли порівнювана величина набагато (у два і більше разів) перевищує базу порівняння, яку беруть за одиницю. *Вони показують, у скільки разів порівнювана величина більша за базу порівняння;*
- **проценти** – застосовують у випадках, коли порівнювана величина не набагато відрізняється від бази порівняння, яку беруть за 100;
- **1‘000** (‰ – **проміле**)**, 10‘000 (**‱ – **продециміле), 100‘000** (‱‰ – **просантіміле)** – застосовують тоді, коли порівнювана величина дуже мала щодо бази порівняння, яку відповідно беруть за 1‘000, 10‘000, 100‘000 (кількість театрів, музеїв бібліотек на 1000 населення, захворюваність, смертність на 100 тис. населення, епідемічний поріг. У 2013 р. кількість людей, яким вперше в житті поставили діагноз онкологічна хвороба 361 на 100‘000 населення, на активний туберкульоз – 68 просантиміле).

Відносні величини можуть бути не лише безрозмірними, але й іменованими, наприклад густота населення – відношення чисельності населення до площі території, на якій воно проживає, ВВП у розрахунку на одну особу. За змістом і відповідно до завдань дослідження виділяють такі види відносних величин:

- **відносні величини планового завдання (ВВПЗ)** характеризують зміни планових (нормативних прогнозних) рівнів, або обсягів, визначених за договорами ($y_{пл}$), порівняно з рівнями, взятими за базові (y_0): $ВВПЗ=y_{пл}/y_0$;
- **відносні величини виконання плану, норми прогнозу (ВВВП)**, що характеризують ступінь виконання плану, визначувані з урахуванням фактичних досягнень (y_1), які порівнюють з рівнями, визначеними планом ($y_{пл}$): $ВВВП=y_1/y_{пл}$;
- **відносні величини динаміки (ВВД)** характеризують напрям та інтенсивність зміни однойменних соціально-економічних явищ у часі і є відношенням рівнів явища за два періоди чи моменти часу: $ВВД=y_1/y_0$.

Зв’язок між наведеними відносними величинами такий: $ВВД=ВВПЗ•ВВВП=(y_п/y_0)•(y_1/y_{пл})=y_1/y_0$.

Всі три види відносних величин можуть бути подані в трьох формах: коефіцієнтів, темпів і темпів приросту у відсотках. У розрахунках (наприклад, для визначення однієї з відносних величин за рештою відомих)

потрібно виразити відносні величини або *у коефіцієнтах,* або у *темпах – з темпами приросту арифметичних дій не виконують.*

Приклад 1.2. *За виробничим планом підприємство мало знизити собівартість продукції за рік на* ***5% (ВВПЗ=0,95)****, а фактично вона знизилась на* ***0,5% (ВВД=0,995)****. Розрахувати ступінь виконання плану:* **ВВД=ВВПЗ•ВВВП**

ВВВП=ВВД/ВВПЗ=0,995/0,95.

ВВВП=ВВД/ВВПЗ=1,0474=104,74%.

Інтерпретуючи отриманий результат, слід мати на увазі, що план із собівартості, трудомісткості, капітало-, ресурсомісткості продукції вважають виконаним, якщо фактичний показник є меншим за плановий, що свідчить про поліпшення роботи підприємства, тоді як зростання наведених показників («місткості») свідчить про *невиконання планового завдання, оскільки досягнута економія ресурсів виявилась меншою, ніж планувалось, а недовиконання плану становить 4,74%.*

Відносні величини структури (ВВС) характеризують склад того чи іншого явища, показуючи, яку частку (питому вагу, %α_i) становлять його окремі складові в усьому явищі. Показники частки обчислюють, ділячи розміри кожної частини сукупності (y_i) на загальну суму частин ($\sum_{i=1}^{n} y_i$):

$$\alpha_i = \frac{y_i}{\sum_{i=1}^{n} y_i}. \qquad (1.5)$$

Зміна в часі питомої ваги окремих частин досліджуваного явища свідчить про **структурні зрушення** (зміни в структурі). У статистиці для оцінювання структурних зрушень використовують два показники:

- **абсолютний приріст** ($\Delta\alpha_i = \alpha_{i1} - \alpha_{i0}$), який обчислюють як різницю між питомою вагою *i*-ї частки досліджуваного явища у звітному та базисному періодах. Абсолютні прирости характеризують швидкість зміни питомої ваги окремих частин досліджуваного явища за розглянутий період, вони завжди різноспрямовані, а тому сума абсолютних приростів структурних зрушень для всіх частин явища завжди дорівнює нулю;
- **коефіцієнт зростання** (зниження) (K_{ai}) є відношенням питомої ваги і-ї частини явища у звітному і базисному періодах: $K_{ai}=a_{i1}/a_{i0}$. Темп зростання структурних зрушень обчислюють, зменшуючи коефіцієнт зростання (зниження) на 1 (100%): $Ta_i=(K_{ai} - 1)$•100%. Відносні показники структурних зрушень відбивають інтенсивність зміни питомої ваги окремих частин досліджуваного явища, точніше, у скільки разів змінилась питома вага певної частки.

Взаємозв'язок між абсолютними та відносними показниками структурних зрушень:

$$\Delta\alpha_i = \alpha_{i1} - \alpha_{i0} = \alpha_{i0}\left(\frac{\alpha_{i1}}{\alpha_{i0}} - \frac{\alpha_{i0}}{\alpha_{i0}}\right) = \alpha_{i0}\left(K_{\alpha_i} - 1\right).$$

Зведеними показниками структурних зрушень, зростання яких свідчить про посилення структурних зрушень є такі:

- **абсолютні показники**;
- **середній лінійний приріст структурних зрушень**, який є середнім арифметичним абсолютних значень абсолютних приростів питомої ваги всіх частин явища, характеризує середню величину відношень питомої ваги, тобто показує, на скільки відсоткових пунктів у середньому відхиляється одна від одної питома вага всіх частин за два порівнюваних періоди:

$$\overline{\alpha} = \frac{\sum_{i=1}^{n}\left|\alpha_{i1} - \alpha_{i0}\right|}{n}, \qquad (1.6)$$

- **середній квадратичний коефіцієнт структурних зрушень** є простим середнім квадратичним абсолютних приростів питомої ваги всіх частин явища за два порівнювані періоди, що показує, на скільки відсоткових пунктів у середньому відхиляється одна від одної питома вага всіх частин за два порівнювані періоди. Цей коефіцієнт має вищу аналітичну цінність, оскільки лінійні показники двох сукупностей можуть збігатись, проте квадратичні коефіцієнти будуть різними**:**

$$\sigma_\alpha = \sqrt{\frac{\sum_{i=1}^{n}\left(\alpha_{i1} - \alpha_{i0}\right)^2}{n}}, \qquad (1.7)$$

- **відносні показники,**
- **лінійний приріст структурних зрушень:**

$$\overline{\alpha} = \sum_{i=1}^{n}\left|\alpha_{i1} - \alpha_{i0}\right|, \qquad (1.8)$$

- **середній квадратичний коефіцієнт структурних зрушень**. Має вищу аналітичну цінність, оскільки лінійні показники двох сукупностей можуть збігатись, проте квадратичні коефіцієнти будуть різними:

$$\sigma_\alpha = \sqrt{\sum_{i=1}^{n}\frac{\left(\alpha_{i1} - \alpha_{i0}\right)^2}{\alpha_{i0}}}. \qquad (1.9)$$

Наведені коефіцієнти мають недолік: вони не дають змоги кількісно оцінити величину інтенсивності зміни структурних зрушень, тому

узагальнювальним показником інтенсивності структурних зрушень, який дає можливість кількісно оцінити розміри структурних змін, є інтегральний коефіцієнт структурних зрушень:

$$K_{\alpha} = \sqrt{\frac{\sum_{i=1}^{n}\left(\alpha_{i1} - \alpha_{i0}\right)^2}{\sum_{i=1}^{n}\alpha_{i1}^{\ 2} + \sum_{i=1}^{n}\alpha_{i0}^{\ 2}}}. \qquad (1.10)$$

Приклад 1.3. *Проаналізувати структурні зрушення у житловому будівництві та будівництві спортивних споруд упродовж t років. Інформаційна база для розрахунків — дані офіційної статистики будівництва, які розміщено на сайті Державної служби статистики України*[1].

Вихідні дані та проміжні розрахунки наведено у табл.1.2, 1.3. У цих таблицях базовий рік позначено як 20xx, а рік, у якому завершено період тривалістю t років, – 20xx+t.

Таблиця 1.2

Вихідні дані та проміжні показники для розрахунку інтегрального коефіцієнта структурних зрушень у житловому будівництві

Показник	Фактичні дані		Розрахунки					
	базовий рік	рік, у якому завершено період тривалістю t років	частка виду будівництва у загальному підсумку		зміна частки	квадрат зміни частки	квадрат частки виду будівництва у загальному підсумку	
	20xx	20xx+t	$\alpha_{i\ 20xx}$	$\alpha_{i\ 20xx+t}$	$\Delta\alpha_i$	$(\Delta\alpha_i)^2$	$(\alpha_{i_{20xx}})^2$	$(\alpha_{i_{20xx+t}})^2$
А	1	2	3= гр1/ підсумок гр.1	4= гр2/ підсумок гр.2	5= гр.4–гр.3	6= (гр.5)^2	7= (гр.3)^2	8= (гр.4)^2
Прийнято в експлуатацію загальної площі житла, тис. кв.м:								
у міських поселеннях	6965	7672	0,740	0,684	–0,056	0,03136	0,5476	0,467856
у сільській місцевості	2445	3545	0,260	0,316	0,056	0,03136	0,0676	0,099856
Разом	**9410**	**11217**	1,000	1,000	**0**	**0,06272**	**0,6152**	**0,567712**

Підставляючи до формули (1.10) результати проміжних розрахунків квадратів часток та їхньої зміни, тобто підсумки граф. 6 – 8 табл. 1.2., одержимо ***значення інтегрального коефіцієнта структурних зрушень для житлового будівництва:***

[1] *http://www.ukrstat.gov.ua*

$$K_{\alpha}^{житл} = \sqrt{\frac{(0{,}74 - 0{,}684)^2 + (0{,}26 - 0{,}316)^2}{0{,}74^2 + 0{,}26^2 + 0{,}684^2 + 0{,}316^2}} = \sqrt{\frac{0{,}006272}{0{,}6152 + 0{,}567712}};$$

$$K_{\alpha}^{житл} = \sqrt{\frac{0{,}006272}{1{,}182912}} = 0{,}2303 = 23{,}03\%.$$

Таблиця 1.3

Вихідні дані та проміжні показники для розрахунку є інтегрального коефіцієнту структурних зрушень у будівництві спортивних споруд

Показник	**Фактичні дані**		**Розрахунки**					
	базовий рік	рік, у якому завершено період тривалістю t років	частка виду будівництва у загальному підсумку		зміна частки	квадрат зміни частки	квадрат частки виду будівництва у загальному підсумку	
	20xx	**20xx+t**	$\alpha_{i\ 20xx}$	$\alpha_{i\ 20xx+t}$	$\Delta\alpha_i$	$(\Delta\alpha_i)^2$	$(\alpha_{i\ 20xx})^2$	$(\alpha_{i\ 20xx+t})^2$
А	**1**	**2**	**3 = гр1/ підсумок гр.1**	**4 = гр2/ підсумок гр.2**	**5 = гр.4– гр.3**	**6 = (гр.5)^2**	**7 = (гр.3)^2**	**8 = (гр.4)^2**
Прийнято в експлуатацію площі спортивних споруд								
Спортивні зали, м^2 тренувальної площі	17573	47083	0,380	0,427	0,047	0,002209	0,1444	0,182329
Плавальні басейни, м^2 дзеркала води	2980	5619	0,065	0,051	–0,014	0,000196	0,004225	0,002601
Площинні спортивні споруди, м^2	25599	57517	0,555	0,522	–0,033	0,001089	0,308025	0,272484
Разом	46152	110219	1,000	1,000	**0**	**0,003494**	**0,45665**	**0,457414**

Аналогічно до попереднього розрахунку у формулу (1.10) підставимо підсумки граф. 6 – 8 табл. 1.3. При цьому ***інтегральний коефіцієнт структурних зрушень для будівництва спортивних споруд*** *становитиме:*

$$K_{\alpha}^{спорт} = \sqrt{\frac{(0{,}427 - 0{,}381)^2 + (0{,}051 - 0{,}065)^2 + (0{,}555 - 0{,}522)^2}{0{,}381^2 + 0{,}065^2 + 0{,}555^2 + 0{,}427^2 + 0{,}051^2 + 0{,}522^2}} = = \sqrt{\frac{0{,}003494}{0{,}45665 + 0{,}457414}};$$

$$K_{\alpha}^{спорт} = \sqrt{\frac{0{,}003494}{0{,}914064}} = 0{,}061826 = 6{,}18\%.$$

Отриманий результат для спортивних споруд виявився майже в чотири рази меншим, аніж для житла: 23,03/6,18=3,72. Це означає, що на тлі ***значних змін у структурі житлового будівництва співвідношення у структурі будівництва спортивних споруд були незначними.***

- **Відносні величини координації (ВВК)** виражають співвідношення, пропорції між окремими частинами явища і визначаються відношенням однієї з його частин до іншої, взятої за базу порівняння. Вони показують, у скільки разів одна частина явища більша за іншу або скільки одиниць однієї частини припадає на 1, 100 і 1000 одиниць іншої частини, взятої за базу порівняння. *Наприклад, у 20хх р. на 1 м^2 житла, зведеного у сільській місцевості, припадало 2,85 м^2 новобудов у міських поселеннях (2,85=6965/2445), в 20хх+t р. це співвідношення скоротилось до 2,16 (=7672/3545).*
- **Відносні величини порівняння (ВВП)** характеризують співвідношення однойменних явищ, або ідентичних показників, що стосуються різних об'єктів (підприємств, галузей), або територій (міст, регіонів, країн) за один і той самий період часу . Базою порівняння може бути будь-який об'єкт, величини виражаються у коефіцієнтах або відсотках, їхня інтерпретація залежить від бази порівняння. *Наприклад, у табл.1.4. наведено вихідні дані для розрахунку відносних величин порівняння.*

Таблиця 1.4

Вихідні дані та проміжні показники для розрахунку відносної величини порівняння економіки країн світу

Показник	Країни		Відносні величина порівняння (ВВП)
	Венесуела	Україна	
А	**1**	**2**	**3=гр.1/гр.2**
ВВП, $ млрд	392	138	2,84
Населення, млн осіб	29	45,4	0,64
ВВП на душу населення $/ос. (р.1/р.2)	**13517**	**3040**	**4,45**

*Згідно з останньою графою табл. 1.4 у 20хх році у **Венесуелі рівень валового виробництва на одну особу був в 4,45 раза вищим, ніж в Україні**. Такий невтішний результат пояснюється тим, що ВВП Венесуели у 2,84 раза перевищував ВВП України, у той час як чисельність населення становила лише 64%, порівняно із нашою державою.*

- **Відносні величини інтенсивності (ВВІ)** характеризують ступінь поширення або розвитку явища в певному середовищі. Вони показують, скільки одиниць однієї сукупності припадає на одиницю іншої: густота населення, народжуваність, смертність, шлюбність, розлучуваність: наприклад, у 20-х роках в Україні народжуваність дорівнювала 11,1, а смертність – 14,6 осіб на 1000 населення.

Забезпеченість населення закладами культури за регіонами (на 100 осіб населення) у 20хх році становила:

– бібліотечний фонд – 685 примірників;

– кількість місць для глядачів у залах для демонстрування фільмів – одне місце;
– кількість місць у клубних закладах – 10 місць.

1.1.4. Середні величини. Математичні властивості середніх величин

Усі середні величини є частковим випадком середньої степеневої:

$$\overline{x_k} = \sqrt[k]{\frac{\sum_{i=1}^{n}(x_i)^k}{n}}. \qquad (1.11)$$

Середня степенева – загальна форма подання різноманітних середніх величин, її записують як $\overline{x_k}$, де *k* – показник степеня, *n* – кількість спостережень, ***i*** — порядковий номер спостереження. Надалі середнє значення будь-якої змінної буде позначене горизонтальною рискою над буквою на позначення такої змінної. За різних значень *k* формула дає різні види середніх:

- ***k=1*** – середня ***арифметична:***

$$\overline{x_{арифм}} = \sqrt[1]{\frac{\sum_{i=1}^{n}(x_i)^1}{n}} \Rightarrow \overline{x_{арифм}} = \frac{\sum_{i=1}^{n} x_i}{n}; \qquad (1.12)$$

- ***k= –1*** – середня ***гармонійна****, яку доцільно використовувати для усереднення відносних величин.* Далі наведено приклад розрахунку цієї середньої:

$$\overline{x_{гарм}} = \sqrt[-1]{\frac{\sum_{i=1}^{n}(x_i)^{-1}}{n}} \Rightarrow \overline{x_{гарм}} = \left(\frac{\sum_{i=1}^{n}\frac{1}{x_i}}{n}\right)^{-1} \Rightarrow \overline{x_{гарм}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}; \qquad (1.13)$$

- ***k=2*** – середня ***квадратична****, яку часто використовують у статистичних та економетричних розрахунках, коли треба визначити мінливість ознаки, що може бути у кожному конкретному спостереженні як більшою, так і меншою за середню:*

$$\overline{x_{квадр}} = \sqrt[2]{\frac{\sum_{i=1}^{n}(x_i)^2}{n}}; \qquad (1.14)$$

- середню ***геометричну*** отримаємо з формули середньої степеневої шляхом *граничного переходу за* $k \to \infty$**:**

$$\overline{x_{геом}} = \sqrt[n]{\prod_{i=1}^{n} x_i}, \qquad (1.15)$$

де $\prod_{i=1}^{n} x_i$ – добуток усіх варіант, тобто значень ознак в досліджуваній сукупності спостережень. Цю середню широко використовують в економічному аналізі, коли визначають усереднені темпи зростання чи приросту певного показника, наприклад, темп інфляції, або зростання доходу, або збільшення років за декілька періодів. Подібним усередненням вимірюють також динаміку курсу національної валюти або цінних паперів на фінансових ринках. Приклад обчислення такої середньої наведено далі. З курсу вищої математики відомо, що **середня геометрична є середньою арифметико-гармонійною двох чисел.** Так, якщо нескінченну кількість разів послідовно усереднювати середні арифметичні та гармонійні двох чисел ***a*** та ***b,*** а також отриманих у подальшому уточнених середніх, одержимо:

$$a_1 = \frac{a+b}{2};\ b_1 = \frac{2}{\frac{1}{a}+\frac{1}{b}} = \frac{2}{\frac{a+b}{a\bullet b}} = \frac{2\bullet a\bullet b}{a+b};$$

$$a_2 = \frac{a_1+b_1}{2};\ b_2 = \frac{2\bullet a_1\bullet b_1}{a_1+b_1}\ \ldots\ a_{n+1} = \frac{a_n+b_n}{2};\ b_{n+1} = \frac{2\bullet a_n\bullet b_n}{a_n+b_n};$$

$$\frac{a+b}{2} > \sqrt{a\bullet b} \Rightarrow \left(\frac{a+b}{2}\right)^2 > a\bullet b \Rightarrow \frac{a+b}{2} > \frac{2\bullet a\bullet b}{a+b}$$

$$\Rightarrow a_n > a_{n+1} > \ldots > \ldots > b_{n+1} > b_{n+1}$$

a та b прямують до спільної межі :

$$a_1\bullet b_1 = a\bullet b \Rightarrow a_n\bullet b_n = a_{n+1}\bullet b_{n+1}\ n\to\infty \Rightarrow c = \sqrt{a\bullet b}.$$

Звичайно, різні види середніх, обчислені за формулами (1.12 – 1.15), мають різні значення, тому застосовують **правило мажоритарності** – **впорядковування середніх величин:**

$$\bar{X}_{\text{гарм}} < \bar{X}_{\text{геом}} < \bar{X}_{\text{арифм}} < \bar{X}_{\text{квадр}}. \qquad (1.16)$$

Різниця між окремими середніми величинами тим більша, чим більшою є варіація усерднюваної ознаки. За незначної варіації така різниця майже непомітна.

Зупинимось на середній **гармонійній**, набагато рідше використовуваній у технічних розрахунках, яка, проте, дає змогу виконувати коректні розрахунки в економічних обґрунтуваннях. Обчислюють середню гармонійну як відношення суми ознак до суми добутків цих ознак на обернені значення варіант. Середня гармонійна може бути не лише простою (1.13), а й зваженою:

$$\overline{x_{гарм}} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{1}{x_i}\bullet w_i}, \qquad (1.17)$$

де w_i – ваговий множник, або вага *i*-го спостереження. Про способи визначення вагових множників йтиметься далі: й у прикладах, і в окремих розділах цього посібника.

Якщо усереднюють лише два спостереження – *a* та *b* з ваговими множниками відповідно w_a та w_b, формула середньої гармонійної зваженої матиме такий вигляд:

$$\overline{x}_{гарм} = \frac{w_a + w_b}{\frac{w_a}{a} + \frac{w_b}{b}} = \frac{(w_a + w_b) \bullet a \bullet b}{w_a \bullet b + w_b \bullet a}. \qquad (1.18)$$

Середню гармонійну застосовують тоді, коли безпосередніх даних про вагу немає, але відомі варіанти усереднюваної ознаки *x* та добутки значень варіант на кількість одиниць, які мають цю ознаку, тобто у випадках, **коли невідомими є абсолютні значення досліджуваних ознак**, наприклад, коли відомими є валові збору урожаю певної культури й величини середньої урожайності на окремих ділянках, однак немає інформації про площі цих ділянок. Якщо є інформація тільки про витрати часу на одиницю продукції та на все виробниче завдання (причому кількість виробів у завданні невідома), то витрати часу на одиничний обсяг робіт в середньому для всієї бригади розраховують за формулою середньої гармонійної.

> ***Приклад 1.4.*** *Проаналізувати продуктивність праці робітників двох конкурентних підприємств, які виконують роботи з улаштування покрівель. Вихідні дані та розрахунки представлено у табл. 1.5. Середні гармонійні обчислюють як зважені за формулою 1.17, вагою, w_i, є показники тривалості роботи на об'єкті.*

Середня гармонійна становитиме:

- ТОВ «Аркадос»:

$$\overline{x}_{гарм}^{Аркадос} = \frac{350 + 480 + 525 + 520}{\frac{350}{5} + \frac{480}{6} + \frac{525}{7} + \frac{520}{8}} = \frac{гр.2\ табл.1.5.}{гр.3\ табл.1.5} = \frac{1875}{290} = 6{,}466\ \text{м}^2/\text{люд год}.$$

- ТОВ «Коймаран»:

$$\overline{x}_{гарм}^{Коймаран} = \frac{600 + 390 + 400 + 490}{\frac{600}{8} + \frac{390}{6} + \frac{400}{5} + \frac{490}{7}} = \frac{гр.5\ табл.1.5.}{гр.6\ табл.1.5} = \frac{1880}{290} = 6{,}483\ \text{м}^2/\text{люд год}.$$

Отже, і продуктивність праці, і коефіцієнт виконання норм виробітку виявляються вищими у підприємства ТОВ «Коймаран»: 6,483>6,466.

Таблиця 1.5

Вихідні дані та розрахунки усереднених показників

ТОВ «Аркадос»				ТОВ «Коймаран»			
робітники	**витрати часу на улаштування 1 м² покрівлі, x_i**	**тривалість роботи на об'єкті, w_i**	*площа облаштованої покрівлі*	**робітники**	**витрати часу на улаштування 1 м² покрівлі, x_i**	**Тривалість роботи на об'єкті, w_i**	*площа облаштованої покрівлі*
А	**1**	**2**	*3=гр.2/гр.1*	**Б**	**4**	**5**	*6=гр.5/гр.4*
1	5	350	*70*	**1**	8	600	*75*
2	6	480	*80*	**2**	6	390	*65*
3	7	525	*75*	**3**	5	400	*80*
4	8	520	*65*	**4**	7	490	*70*
Разом		**1875**	**290**	**Разом**		**1880**	**290**

Загалом, якщо показники випуску продукції окремих підприємств та показники виробітку на працівника відомі, проте немає інформації про чисельність працівників, краще середній виробіток обчислити як середню гармонійну. Те саме стосується й аналізу ефективності роботи персоналу на підприємствах торгівлі, й усереднення характеристик ринку нерухомості.

Приклад 1.5. *Відомо, що за останній місяць продано три об'єкти нерухомості за 81, 120 та 130 млн грн, а також той факт, що коефіцієнти капіталізації цих об'єктів (тобто співвідношення ціни продажу до потенційного валового доходу, що визначається корисною площею об'єкта, яка може бути здана в оренду, й орендної ставки) відповідно становлять 1,35; 1,5 й 1,3.* ***Усереднений коефіцієнт капіталізації*** *дорівнюватиме:*

$$\bar{k}_h = \frac{\sum_{i=1}^{n} z_i}{\sum_{i=1}^{n} \frac{z_i}{k_i}} = \frac{81+120+130}{\frac{81}{1{,}35}+\frac{120}{1{,}5}+\frac{130}{1{,}3}} = \frac{81+120+130}{60+80+100} = \frac{331}{240} \approx 1{,}38;$$

$$\bar{k}_h = \frac{\sum_{i=1}^{n} \alpha_i}{\sum_{i=1}^{n} \frac{\alpha_i}{k_i}} = \frac{0{,}24+0{,}36+0{,}40}{\frac{0{,}24}{1{,}35}+\frac{0{,}36}{1{,}5}+\frac{0{,}40}{1{,}3}} = \frac{1}{0{,}18+0{,}240+0{,}30} = \frac{1}{0725} \approx 1{,}38.$$

Розрахунок середньої гармонійної є корисним для планування й організації будівельного виробництва, а також інших видів комерційної діяльності у будівництві, коли визначальним є своєчасне поповнення запасів.

***Приклад 1.6**. Автоцементовоз їхав від виробника цементу до об'єкта будівництва 40 хв, а повертався для наступного навантаження на заводі за 35 хвилин. Причому в обох поїздках довжина маршруту була однаковою – 38 км. Визначаємо середній час поїздки:*

$$\bar{t}_h = \frac{S}{V} = \frac{\sum_{i=1}^{n} S_i}{\sum_{i=1}^{n} \frac{S_i}{t_i}} = \frac{2 \cdot 38}{38 \cdot \frac{60}{35} + 38 \cdot \frac{60}{40}} = \frac{76}{65{,}14 + 57} = 0{,}6(2) \text{ год або}$$

$0{,}622 \cdot 60 \approx 37{,}(3)$ *хв*.

У статистичних розрахунках відомий і такий вид середньої, як середня антигармонійна.

Середня антигармонійна – це відношення суми других степенів (квадратів) індивідуальних значень *усереднюваної ознаки* до суми їхніх перших степенів, яке обчислюють за формулою: $\bar{x} = \frac{\sum x^2}{\sum x}$, де *x* – індивідуальні значення усередненої ознаки.

Наведемо приклад розрахунку середньої гармонійної та її економічної інтерпретації.

***Приклад 1.7**. У табл. 1.7 наведено інформацію про ціну керамзиту в м. Києві у 2016 – 2021 рр. Джерело даних – інформація про ціни будівельних матеріалів, наведена на сайті науково-виробничої фірми «Інпроєкт»*[2]. За базовий рік взято 2016-й, а тому порядковий номер відповідного спостереження *i*=0 (а не 1). Останнє спостереження відповідне 2021 року і, таким чином, має номер 5: *i*=5. В аналізі часових рядів економічних даних важливе місце належить усередненому темпу зростання. В розглядуваному прикладі це – темп зростання ціни керамзиту. Для його обчислення виконують усереднення ланцюгових темпів підвищення ціни за формулою (1.15). Ланцюгові темпи зростання ціни визначають, порівнюючи ціни поточного та попереднього періодів. У наведеному прикладі

періоди – це роки. Розрахунок ланцюгових темпів зростання виконано в графі 2 табл. 1.7. Для базового року беруть ланцюговий індекс, рівний 1. Його, таким чином, не варто враховувати, усереднюючи за формулою (1.15). Цим пояснюється нульовий індекс на позначення номера спостереження та величина кількості використовуваних в розрахунках спостережень *n*, на 1 менша за

[2] https://www.inproekt.kiev.ua/CO/Advice

фактичну кількість зведених у табл. 1.7 періодів. В останньому рядку графи 2 табл. 1.7 наведено базовий індекс цін. Це співвідношення останнього рівня ціни до базового.

Таблиця 1.7

Вихідні дані та проміжні розрахунки для визначення середньорічного зростання ціни керамзиту в м. Києві

Період	Порядковий номер спостереження, і	Ціна керамзиту, грн/м^2	Ланцюговий темп зростання ціни керамзиту (ланцюговий індекс цін)	
А	Б	1	2=гр.1/попереднє значення гр.1	
			значення	розрахунок
2016	0	800	1	
2017	1	900	1,125	900/800
2018	2	1000	1,111	1000/900
2019	3	1100	1,110	1100/1000
2020	4	1200	1,091	1200/1100
2021	5	1300	1,083	1300/1200
Базовий індекс цін			1,625	1300/800

Базовий індекс ціни керамзиту у 2021 р. відносно 2016 р. становив **1,625**. Це означає, що упродовж останніх п'ятьох років ціни зросли в 1,625 раза, або на 62,5%. Проте зростання цін відбувалось з різним темпом, адже щорічні

100-гривневі збільшення ціни (зміна значень в р.1) порівнюють із різною базою попереднього року: у 2017 р. 100-гривневе підвищення від 800 грн становило 12,5% приросту ціни (індекс 1,25), тоді як 100 грн, на які зросла ціна керамзиту у 2020 р. проти 2019 р., коли вона становила 1100 грн/м^3, є відповідним індексу 1,091, чи 9,1%-му приросту. Усереднювати значення індексів ціни за формулою простої середньої арифметичної (1.12) не слід, оскільки розрахунки будуть виконані над **відносними величинами**.

Оскільки в розрахунку середньої арифметичної операції додавання виконують над первинними даними, усереднюючи темпи зростання (тобто індекси), доцільно виконувати арифметичні дії подібного порядку. Для відносних величин – індексів, отриманих як результат ділення первинних даних, доцільно використати не додавання, а множення. Тоді наступна дія має бути також вищого порядку: не ділення суми, а визначення кореня з добутку. Таким чином, для того щоби правильно визначити середній темп зростання ціни (та й будь-якого іншого показника), потрібно розрахувати середню геометричну (1.15). Крім того, відповідно до правила мажоритарності (1.16) отримана середня не буде завищувати темпи зростання, порівняно з середньою арифметичною.

Отже, середньорічне зростання цін на керамзит у м. Києві у 2016 – 2021 рр. дорівнюватиме:

$$\overline{x_{геом}} = \sqrt[5]{\frac{900}{800} \cdot \frac{1000}{900} \cdot \frac{1100}{1000} \cdot \frac{1200}{1100} \cdot \frac{1300}{1200}} = \sqrt[5]{1{,}125 \cdot 1{,}111 \cdot 1{,}110 \cdot 1{,}091 \cdot 1{,}083} = \\ = \sqrt[5]{1{,}625} = 1{,}102. \quad (1.19)$$

Таким чином, щороку ціна керамзиту зростає в середньому на 10,2%, або середньорічний темп зростання ціни становить 1,102 раза. Цю обставину слід мати на увазі під час складання річних бюджетів доходів, матеріальних витрат, грошових надходжень та витрат на підприємствах будівництва.

Із розрахунку (1.19) помітно, що формулу для розрахунку середнього темпу зростання можна значно спростити, скоротивши складові обчислення ланцюгових темпів зростання. Таким чином, **середньорічний темп зростання** – корінь, степінь якого на 1 менший за кількість вихідних даних, із базового темпу зростання:

$$\overline{Тзрост} = \sqrt[n]{\frac{x_n}{x_0}} = \sqrt[n]{Тзрост_{баз}}. \quad (1.20)$$

Якщо базовий період позначити не як «0», а звичайно як «1», формула дещо зміниться – буде інший показник степеня кореня. Крім того, наголосимо, що в часових рядах номер спостереження позначають не буквою «*i*», а буквою (*t*):

$$\overline{Тзрост} = \sqrt[t-1]{\frac{x_t}{x_1}} = \sqrt[t-1]{Тзрост_{баз}}. \quad (1.21)$$

У практиці статистичного аналізу найчастіше доводиться обчислювати такі різновиди середньої **арифметичної**:

- **середню з групових середніх**:

$$\overline{x} = \frac{\sum_{j=1}^{m} \overline{x}_j \cdot f_j}{\sum_{j=1}^{m} f_j}, \quad (1.21)$$

де $\overline{x}_j$ – середнє значення ознаки в групі під номером *j*; f_i – обсяг *j*-ї групи; *m* – кількість груп.

- **середню величину альтернативної ознаки**, варіація якої виражається двома значеннями: 1 – за наявності ознаки або 0 – за відсутності, можна розрахувати, позначивши частку одиниць, що мають ознаку, як *p*, а частку одиниць, що не мають цієї ознаки, як *q*=(1 – *p*). Тоді:

$$\overline{x}_{альт} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} = \frac{p \cdot 1 + q \cdot 0}{p + q} = \frac{p \cdot 1}{1} = p. \quad (1.22)$$

Приклад 1.8. *З 50 відвідувачів магазину освітлювальної техніки покупку зробили 42. Визначити середній коефіцієнт здійснення покупок:* ***p=42/50=0,84***;

q=1 - 0,84=0,16;

k=(0,84•1+0,16•0)/(0,84+0,16)=0,84.

Середні величини мають такі математичні властивості:

- середня арифметична постійної величини дорівнює їй самій: $\overline{A}=A$;
- постійний множник може бути винесеним за знак середньої арифметичної:

$$\overline{\mathrm{A}} \cdot \bar{x} = A \cdot \bar{x};$$

добуток кількості одиниць сукупності та середньої арифметичної дорівнює сумі значень усередненої ознаки, тобто визначальному показнику: $n \bullet \overline{x} = \sum_{i=1}^{n} x_i$ та $\overline{x} \bullet \sum_{i=1}^{n} f_i = \sum_{i=1}^{n} x_i \bullet f_i$;

- алгебраїчна сума відхилень усіх варіант від середньої арифметичної дорівнює нулю:

$$\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = 0;\ \sum_{i=1}^{n} x_i - n \bullet \overline{x} = \sum_{i=1}^{n} x_i - n \bullet \frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0;$$

$$\sum_{i=1}^{n}\left(x_i - \overline{x}\right) \bullet f_i = 0.$$

Саме через цю властивість підсумок графи 5 у табл. 1.2 та 1.3 дорівнював нулеві;

- сума квадратів відхилень окремих варіант ознаки від середньої арифметичної є меншою, ніж від будь-якої іншої величини, тобто

$$\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \bullet f_i > \sum_{i=1}^{n}\left(x_i - A\right)^2 \bullet f_i, \text{ або } \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \bullet f_i = \min ;$$

- якщо всі варіанти збільшити або зменшити на будь-яку величину, то середня арифметична зміниться відповідно на ту саму величину:

$$\frac{\sum_{i=1}^{n}\left(x_i \pm A\right) \bullet f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i \bullet f_i}{\sum_{i=1}^{n} f_i} \pm \frac{\sum_{i=1}^{n} A \bullet f_i}{\sum_{i=1}^{n} f_i} = \overline{x} + A ;$$

- якщо всі варіанти поділити чи помножити на будь-яке довільне число *A*, то середня арифметична зменшиться або збільшиться в *A* разів:

$$\frac{\sum_{i=1}^{n}(x_i / A)\cdot f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} / \frac{\sum_{i=1}^{n} A \cdot f_i}{\sum_{i=1}^{n} f_i} = \bar{x} / A;$$

$$\frac{\sum_{i=1}^{n}(x_i \cdot A)\cdot f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} \cdot \frac{\sum_{i=1}^{n} A \cdot f_i}{\sum_{i=1}^{n} f_i} = \bar{x} \cdot A;$$

- якщо ВСІ ЧАСТОТИ поділити чи помножити на будь-яке довільне число А, то середня арифметична від цього не зміниться:

$$\frac{\sum_{i=1}^{n}(f_i / A)\cdot x_i}{\sum_{i=1}^{n} f_i / A} = \frac{A \cdot \sum_{i=1}^{n} x_i \cdot f_i}{A \cdot \sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} = \bar{x};$$

$$\frac{\sum_{i=1}^{n}(f_i \cdot A)\cdot x_i}{\sum_{i=1}^{n} f_i \cdot A} = \frac{A \cdot \sum_{i=1}^{n} x_i \cdot f_i}{A \cdot \sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} = \bar{x}.$$

Ця властивість дає змогу дійти висновку, що величина середньої арифметичної залежить не від абсолютних значень частот (ваг), а й від їхнього співвідношення в сукупності, тобто від частки кожної варіанти у сукупності, а тому замість $\frac{f_i}{\sum_{i=1}^{n} f_i}$ можна вживати α_i:

$$\bar{x} = \sum_{i=1}^{n} x_i \cdot \alpha_i, \ \sum_{i=1}^{n} \alpha_i = 1.$$

Середня гармонійна зважена, як і середня арифметична зважена, не зміниться, якщо визначальний показник*, записаний в чисельнику формули розрахунку і який є вагою окремих варіант, помножити або поділити на будь-яке довільне число, тому середню гармонійну зважену можна обчислювати не лише на основі абсолютних величин, але і їхньої питомої ваги:*

$$\bar{x}_h = \frac{\sum_{i=1}^{n} \alpha_i}{\sum_{i=1}^{n} \frac{\alpha_i}{x_i}}, \ \alpha_i = \frac{n_i}{\sum_{i=1}^{n} n_i}.$$

1.2. СТАТИСТИЧНЕ СПОСТЕРЕЖЕННЯ, ЗВЕДЕННЯ ТА ГРУПУВАННЯ

1.2.1. Основні теоретичні відомості

Статистичне спостереження – науково організований збір масових даних про явища та процеси, що відбуваються в суспільстві. Слід відрізняти *одиницю статистичного спостереження* від *елемента сукупності.* Одиниця – носій інформації, елемент – носій ознак: під час перепису машин й устаткування одиницею спостереження є окреме підприємство, а елементом – окрема машина чи механізм. Застосовують *дві організаційні форми спостереження – звітність і спеціально організовані спостереження,* що охоплюють переписи, одноразові обліки, вибіркові обстеження. Види спостережень систематизовано (рис.1.2).

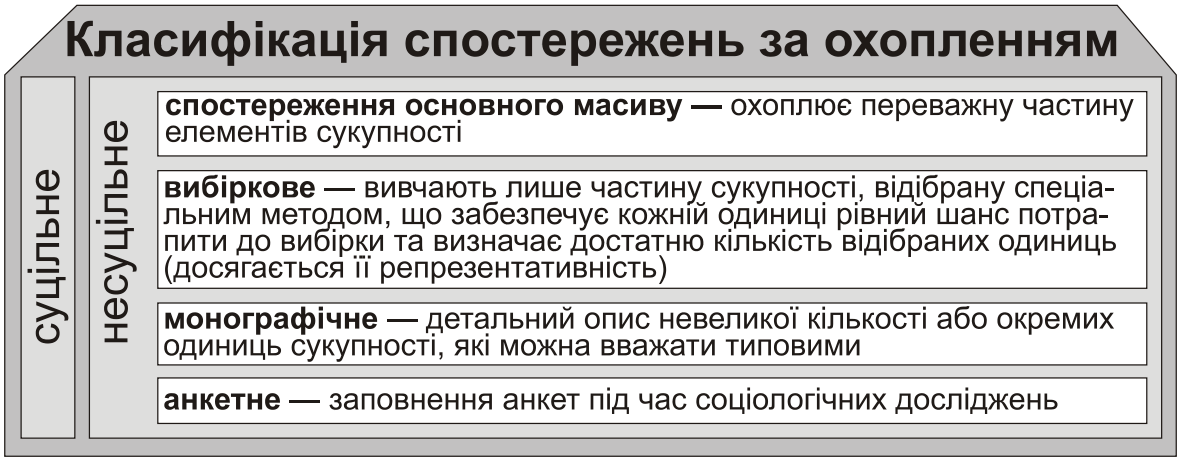


Рис. 1.2. Класифікація статистичних спостережень

Помилки спостережень — розбіжності між даними спостережень та реальними даними. Розрізняють:

- помилки реєстрації, які виникають внаслідок неправильного встановлення фактів, або реєстрації;
- помилки репрезентативності, властиві тільки вибірковому спостереженню, виникають внаслідок некоректного формування вибіркової сукупності.

Обидва види помилок можуть бути як систематичними, так і випадковими (несистематичними):

- випадкові помилки реєстрації – це наслідок впливу різних випадкових чинників (наприклад, механічні під час заповнення формуляра), вони можуть як завищувати, так і занижувати значення показників. За достатньо великої сукупності ці помилки згідно із законом великих чисел взаємно нейтралізуються;
    - систематичні помилки реєстрації завжди мають однакову тенденцію або лише завищення, або лише заниження значення показників за кожною одиницею спостереження, тому величина показника сукупності буде містити накопичену помилку;

- випадкові помилки репрезентативності виникають, якщо відібрана (вибіркова) сукупність не цілком відтворює генеральну;
- систематичні помилки репрезентативності з'являються, якщо порушено принципи відбору одиниць з генеральної сукупності.

Арифметичний *та логічний контроль потрібні для своєчасного виявлення помилок: помноживши якісний показник на кількісний, маємо одержати загальний: інформація про чисельність працівників та їхній середній виробіток (зарплату) дає змогу перевірити правильність визначення загального обсягу випуску продукції (чи фонду оплати праці).* ***Логічний*** *контроль полягає у такому: у збиткового підприємства не може бути додатних показників рентабельності.*

1.2.2. Поняття про зведення та групування. Види групувань

У результаті статистичного спостереження постає потреба абстрагуватись від несуттєвого та випадкового й виокремити найважливіше, тобто в спеціальному обробленні статистичних даних – ***зведенні матеріалів спостереження.***

Зведення – комплекс дій з узагальнення конкретних індивідуальних даних, які утворюють сукупність, з метою виявлення типових рис і закономірностей, властивих досліджуваному явищу загалом. Зведення охоплює:

- розробкулення або вибір показників, які характеризують типові групи та підгрупи;
- групування матеріалу;
- підрахунок групових та загальних підсумків;
- викладення результатів у вигляді статистичних таблиць і графіків.

Статистичне групування – розподіл сукупності на групи за істотними для них ознаками. За допомогою статистичного групування вирішують три основних завдання, що визначає основні типи групувань**:**

- поділення неоднорідної сукупності на якісно однорідні групи за допомогою *типологічних групувань* (поділ працівників за рівнем освіти або обсяги виконаних будівельних робіт за видами продукції);
- вивчення *структури* та структурних зрушень в якісно однорідних сукупностях, їхній розподіл за величиною варіювальної ознаки (поділ працівників за віком або розподіл кількості незавершених будівель та інженерних споруд, які перебувають у стадії будівництва, за рівнем будівельної готовності);
- виявлення та вивчення взаємозв'язку між факторними й результативними ознаками виконують за допомогою *аналітичних групувань* (поділ працівників за рівнем кваліфікації із зазначенням

середньомісячної зарплати для кожної групи, розподіл квартир за кількістю кімнат та загальною площею).

Факторні ознаки є причиною інших ознак і зумовлюють їхні зміни, **результативні ознаки** – є наслідками дій факторних ознак і змінюються під їхнім впливом: ***y=f(x)*** – *результат y – функція фактора x.*

У процесі побудови аналітичних групувань сукупності поділяють на групи за факторною ознакою, в кожній групі розраховують середнє значення результативної ознаки. Якщо в міру зростання значень факторної ознаки систематично зростає чи зменшується середнє значення результативної ознаки, то це свідчить про наявність прямого чи зворотного зв'язку між цими ознаками. Брак будь-якої систематичності у зміні середніх значень результативної ознаки свідчить про брак зв'язку.

Під час виконання будь-якого групування розв'язують три питання:

- ***вибір групувальної ознаки;***
- ***визначення кількості груп;***
- ***визначення меж, за якими відокремлюються окремі групи.***

Інтервал – це деяка зона варіювання кількості ознаки, обмежена відповідно ліворуч і праворуч двома певними значеннями (межами). Найменше значення ознаки, яке позначає інтервал, називається ***нижньою межею*** інтервалу, а найбільше – ***верхньою межею***. Зазвичай нижню межу вважають «включною», а верхню – «виключною», тобто інтервалом [)

Розмір, або величина, ***інтервалу*** (***h***), – це різниця між межами інтервалу, $h=(x_{max} - x_{min})/m$ (де m – кількість груп). Інтервали можуть бути не лише однаковими, а й різними за розміром:

- **спеціалізовані** – це такі нерівні інтервали, які застосовують для розподілу за однією й тією самою ознакою одиниць різних сукупностей для виділення у них однакових типів, груп;
- **прогресивного зростання** або **спадні інтервали** – це такі нерівні інтервали, величина яких збільшується або зменшується на основі прогресій: *арифметичної* ($h_{i+1}=h_i+d$) або *геометричної* ($h_{i+1}=h_i \cdot q$), де d, q – *константи відповідних прогресій*;
- **довільні інтервали** – це такі нерівні інтервали, величина яких залежить від характеру варіації групувальної ознаки і може бути визначена методом рівних частот.

За наявністю меж інтервали поділяють так:

- **закриті** – інтервали, які мають позначення нижньої та верхньої меж;
- **відкриті** – інтервали, в яких немає позначень нижньої та верхньої меж: перший інтервал містить слова «до» або «менше», а останній – «понад» або «більше».

Величина інтервалу залежить від кількості інтервалів і варіації досліджуваної ознаки. Що ширший розмір коливання $x_{max} - x_{min}$, то більшою буде величина інтервалу h, що більше m, то менша величина інтервалу.

Завеликі інтервали викривлюють сутність досліджуваних явищ, а замалі призводять до необґрунтовано нечисленних груп.

Орієнтовно кількість інтервалів (груп) можна визначити за формулою, рекомендованою у 1926 р. американським статистиком Х. Стерджесом (Sturgess):

$$m=1+3,322 \cdot lg(n) = 1+lg(n)/lg(2) \quad (1.23)$$

або

$$m=1+1,441 \cdot ln(n) = 1+ln(n)/ln(2) \quad (1.24)$$

де m – кількість інтервалів;

n – кількість одиниць сукупності.

Результат формули заокруглюють зазвичай до більшого числа (табл.1.8).

Таблиця 1.8

Орієнтовна кількість груп залежно від кількості спостережень за формулою (1.24)

n	**2**	**3**	**4**	**5**	**6**
ln	0,693	1,099	1,386	1,609	1,792
кількість груп	1,443	2,586	3,000	3,322	3,586
	1	3	3	3	4

n	**7**	**8**	**9**	**10**	**11**	**12**	**13**
ln	1,946	2,079	2,197	2,303	2,398	2,485	2,565
кількість груп	3,808	4,000	4,170	4,323	4,460	4,586	4,701
	4	4	4	4	4	5	5

n	**14**	**15**	**24**	**25**	**44**	**45**	**89**	**90**
ln	2,639	2,708	3,178	3,219	3,784	3,807	4,489	4,500
кількість груп	4,808	4,908	5,586	5,645	6,460	6,494	7,478	7,494
	5	5	6	6	6	6	7	7

n	**179**	**180**	**359**	**360**	**719**	**720**	**1439**
ln	5,187	5,193	5,883	5,886	6,578	6,579	7,272
кількість груп	8,485	8,494	9,489	9,494	10,492	10,494	11,494
	8	8	9	9	10	10	11

З формули Стерджеса й табл. 1.8. випливає, що кількість груп визначається обсягом сукупності – вона зростає в міру збільшення кількості спостережень, як показано у табл. 1.9.

Таблиця 1.9

Зв’язок між кількістю груп та кількістю спостережень за формулою (1.24)

Кількість спостережень, n	3-5	6-11	12-24	25-44	45-94	95-194	195-384	385-774	775-1556
Кількість груп, m	3	4	5	6	7	8	9	10	11

У деяких випадках здійснюють перегрупування даних з метою утворення нових груп на основі наявних, якщо останні не задовольняють мету

аналізу. У такий спосіб утворюються *вторинні групування,* отримані на основі перегрупувань *первинних. Вторинні групування можна одержати 3-ма способами:*

- перегрупування однієї із сукупностей за інтервалами іншої або перегрупування обох сукупностей за новими інтервалами;
- за питомою вагою груп у їхньому загальному підсумку;
- перехід від структурного групування, побудованого за кількісною ознакою, до типологічного.

Приклад 1.9. Виконаємо статистичне групування для обсягів попиту на автомобілі у світі. Вихідні дані та результати групувань за відповідними сукупностями представлено у табл. 1.10, Таблиця містить упорядковані дані річного продажу автомобілів. Оскільки країни впорядковано за убуванням кількості проданих машин, порядкові номери груп також убувають.

Таблиця 1.10

Групування **обсягів продажів автомобілів у 2019 році, шт.**
(за даними світового атласу статистичних даних [3])

Країна	к-ть авто	група	Країна	к-ть авто	група
Венесуела	4500	6	Куба	3015	4
Монголія	3800	5	Ангола	2850	4
Албанія	3400	4	Уганда	2500	3
Грузія	3400	4	Габон	2461	3
Замбія	3255	4	Таджикистан	2450	3
Камерун	3079	4	Зімбабве	2150	3
Мадагаскар	1900	2	Киргизія	1505	2
Танзанія	1900	2	Вірменія	1210	1
Судан	1696	2	Йемен	1200	1
Багами	1644	2	Конго	1089	1
Нікарагуа	1600	2	Буркіна Фасо	1074	1
Малаві	1538	2	Ботсвана	1050	1

- **Кількість груп**, **k=1+ln(24)/ln(2) ≈ 6.**
- **Крок** інтервалу, ***h***, зважаючи на те, що **кількість внутрішніх меж груп завжди на одну менша за кількість груп:**

$$h = \frac{X_{max} - X_{\min}}{k - 1} = \frac{4500 - 1050}{6 - 1} = 690.$$

- **Межі груп** визначають так, щоб мінімальне і максимальні значення потрапили в середину інтервалу. Тому для першої групи ліва й права

[3] https://knoema.ru/atlas/topics/%d0%a2%d1%80%d0%b0%d0%bd%d1%81%d0%bf%d0%be%d1%80%d1%82/Motor-Vehicle-Sales/Car-sales

межі визначають відповідно додаванням і відніманням від мінімального значення пів ширини інтервалу:

1050+690•0,5=1095;

1050—690•0,5=705.

Тоді межі подальших інтервалів будуть розміщені так, щоби максимум і мінімум обстежуваної сукупності були в центрі крайніх груп. В цьому прикладі **межі** інтервалів і частоти груп f_j, наведено в дужках**:**

№1: 706—1395 (f_1=5);

№2: 1396—2085 (f_2=7)**;**

№3: 2086—2775 (f_3=4)**;**

№4: 2776—3465 (f_4=6)**;**

№5: 3466—4155 (f_5=1)**;**

№6: 4156—4845(f_6=1).

З огляду на те, що частоти груп № 5 та № 6 дорівнюють 1, ці дві групи доцільно об'єднати.

1.3. МЕТОДИ АНАЛІЗУ РЯДІВ РОЗПОДІЛУ

1.3.1. Поняття про ряди розподілу та їхні основні складові

Одиниці статистичної сукупності характеризуються багатьма ознаками, значення яких змінюється від однієї одиниці до іншої. Для того щоби визначити характер розподілу одиниць досліджуваної сукупності за варіаційними ознаками, отримати найповнішу характеристику її складу, виявити закономірності розподілу її одиниць за тією чи іншою ознакою, будують ряди розподілу за будь-якою ознакою.

Ряд розподілу – це такий розподіл одиниць статистичної сукупності за значенням будь-якої ознаки, коли кожному значенню або групі значень цієї ознаки відповідною є деяка кількість одиниць сукупності.

У загальному вигляді будь-який ряд розподілу є таблицею, що складається з двох рядків (стовпчиків), в одному з яких містяться різновиди чи окремі значення досліджуваної ознаки – варіанти, x_i, в другому – абсолютні числа, що показують, скільки разів повторюються окремі значення варіанти – частоти, f_i. ***Сума всіх частот*** ряду розподілу, що дорівнює загальній чисельності одиниць досліджуваної сукупності, називається ***об'ємом ряду розподілу***, або ***обсягом сукупності***, який позначають як n:

$$n = f_1 + f_2 + f_3 + \ldots + f_i + \ldots + f_n = \sum_{i=1}^{n} f_i \,. \qquad (1.25)$$

Між групуванням та рядами розподілу є певна подібність, оскільки ряди розподілу, як і групування, утворюють внаслідок групування одиниць

статистичної сукупності за будь-якою ознакою, проте статистичне групування є первісним етапом оброблення статистичних даних, тоді як ряд розподілу дає змогу більш глибоко вивчити закони розподілу ознак, оцінити імовірність появи тієї чи іншої варіанти у подальших спостереженнях.

Характеристиками рядів розподілу є частоти, частки, кумулятивні частоти і частки.

Частоти (f_i) характеризують ступінь поширеності варіанти або групи варіант серед одиниць сукупності.

Частки (α_i)**,** тобто модифіковані частоти, визначають діленням окремих варіант на загальну кількість одиниць сукупності:

$$\alpha_i = \frac{f_i}{\sum_{i=1}^{n} f_i} = \frac{f_i}{n}. \qquad (1.26)$$

Сума всіх часток дорівнює одиниці:

$$1 = \frac{n}{n} = \frac{f_1}{n} + \frac{f_2}{n} + \frac{f_3}{n} + \ldots + \frac{f_i}{n} + \ldots + \frac{f_n}{n} = \sum_{i=1}^{n} \alpha_i ,$$

*за умови, що частки виражені у коефіцієнтах***.** Проте частки, виражені у коефіцієнтах , досить важко сприймаються у практичній роботі, тому їх подають у зручнішому для сприйняття та аналізу вигляді, помноживши на 100 (% – процент), 1‘000 (‰ – проміле), 10‘000 (%оо – продециміле), 100‘000 (%ооо – просантіміле). У такому разі сума всіх часток ряду дорівнюватиме відповідно 100, 1‘000, 10‘000, 100‘000.

Кумулятивні частоти (S_{fi}) і частки (S_{ai}) обчислюють, підсумовуючи частоти або частки такої варіанти чи групи варіант ряду розподілу частот або часток всіх попередніх варіант, *тобто кумуляцією частот або часток зверху вниз, починаючи з частоти або частки першої варіанти.* Для першої варіанти її кумулятивна частота (частка) збігається із її власною часткою (частотою). Для останньої варіанти кумулятивна частота збігається з обсягом сукупності (одиницею для коефіцієнтів, 100 для відсотків тощо).

Кумулятивні частоти (S_{fi}) означають, скільки одиниць сукупності не перевищують такого значення ознаки, тобто мають варіанту, не вищу за дану.

Кумулятивні частки (S_{ai}) показують, яка частка (відсоток) одиниць сукупності не перевищує дану варіанту.

Частоти і частки дають змогу дослідити характер розподілу одиниць сукупності за певною ознакою, форму цього розподілу і ступінь коливання, мінливості значень такої ознаки. Кумулятивні частки і частоти використовують для побудови *кумулятивних рядів*, за допомогою яких можна не лише отримувати важливі показники для характеристики структури досліджуваної сукупності, а й вивчати процеси концентрації та диференціації досліджуваного явища, ступінь його диференціації, визначати важливі характеристики ряду розподілу – *порядкові статистики – квантилі*, що поділяють сукупність на ряд однакових за чисельністю частин.

Кількісні ряди розподілу, залежно від способу побудови бувають таких видів:

- **дискретні** – побудовані на основі конкретних значень кількісних ознак, і перервних, і неперервних, наприклад, розподіл домогосподарств за розміром. Їх будують, коли кількісна ознака набуває малу кількість значень – до 10 – 15. При цьому питання про кількість груп не виникає – їх стільки ж, скільки й ознак;
- **інтервальні** – у разі широких меж вирівнювання ознаки (як дискретної, так і неперервної) для їхньої побудови вдаються до тих самих методів, що й під час групування*:*
  - *рівних інтервалів;*
  - *рівних часток;*
  - *кратності інтервалів.*

В інтервальних рядах розподілу з нерівномірними інтервалами безпосередньо порівнювати частоти не можна, оскільки їхні величини залежать не лише від значень ознаки, які зумовлюють межі інтервалів, а й від величини інтервалів: *що ширший інтервал, то більше він матиме одиниць сукупності*.

Для того щоби забезпечити порівнюваність частот або часток рядів розподілу з нерівномірними інтервалами, обчислюють щільність інтервалів, або розподілу, яка показує, скільки одиниць сукупності або який їхній відсоток припадає в середньому на одну одиницю величини інтервалу. Щільність розподілу визначають як результат ділення частоти або частки (f_i, a_i) на його величину (h_i):

- **абсолютна щільність розподілу:** $D_a = \frac{f_i}{h_i}$ **;**
- **відносна щільність розподілу:** $D_в = \frac{\alpha_i}{h_i}$ **;**

Для статистичного вивчення рядів розподілу використовують чотири групи статистичних показників:

- **характеристики центру розподілу і порядкові статистики;**
- **характеристики варіації;**
- **характеристики форми розподілу;**
- **характеристики диференціації та концентрації.** Показники цієї групи більшою мірою стосуються макроекономічного аналізу рівномірності розподілу доходів у суспільстві.

Перші дві групи показників надзвичайно важливі для статистичного аналізу даних, тому розглянемо їх докладніше. Характеристики форми розподілу дають змогу дуже наближено з'ясувати, якою мірою розподіл ознаки в аналізованій сукупності наближається до нормального, однак розрахункові процедури цих показників вельми трудомісткі. Більш точний

спосіб перевірки відповідності розподілу ознаки нормальному розподілу описано у розділі посібника, у якому висвітлено перевірку статистичних гіпотез.

Наведемо приклад складання ряду розподілу, результати якого будуть використані під час розрахунку основних груп статистичних показників.

Приклад 1.10. Потрібно скласти ряд розподілу угод щодо оренди торговельних приміщень. Вихідні дані та результати групувань представлено у табл. 1.11. У таблиці різні групи виділено кольором та форматуванням. Через те що дані розміщено в хронологічній послідовності, процес групування є більш трудомістким.

Таблиця 1.11

Групування кількості угод оренди торговельних приміщень у липні – серпні 20-го року

Дата	Кількість укладених угод	Група	Дата	Кількість укладених угод	Група	Дата	Кількість укладених угод	Група
20.7	*192*	*3*	***3.8***	***174***	***2***	17.8	275	5
21.7	286	6	**4.8**	**159**	**1**	18.8	287	6
22.7	286	6	5.8	288	6	19.8	249	4
23.7	281	5	**6.8**	**159**	**1**	20.8	225	4
24.7	296	6	***7.8***	***180***	***2***	21.8	246	4
27.7	275	5	*10.8*	*202*	*3*	25.8	242	4
28.7	286	6	11.8	269	5	**26.8**	**156**	**1**
29.7	284	6	12.8	290	6	27.8	257	5
30.7	*213*	*3*	13.8	233	4	28.8	269	5
31.7	*201*	*3*	***14.8***	***182***	***2***	**31.8**	**146**	**1**

Кількість груп, ***k=1+ln(30)/ln(2)≈6,*** **крок** інтервалу $h = \frac{296 - 146}{6 - 1} = 30$**,**

межі інтервалів**:**

№1 до 160 *(≈7,2+4,3•0,5),* f_1*=4 дні;*

№4: 221—250 f_4*=5 днів***;**

№2: 161—190 f_2*=3 дні***;**

№5: 251—280 f_5*=5 днів***;**

№3: 191—220 f_3*=4 дні***;**

№6: понад 280 f_6*=9 днів***.**

Результати обчислень зведемо у табл. 1.12, яка, власне, і є рядом розподілу.

Таблиця 1.12

Ряд розподілу кількості угод оренди торговельних приміщень у липні – серпні 20-го року

Група	Межі групи	Частота, f_j	Кумулятивна частота, S_{fj}	Частка, w_j	Кумулятивна частка, S_{wj}
1	до 160	4	4	0.13	0.13
2	161 – 190	3	7	0.10	**0.23**
3	191 – 220	4	11	0.13	0.37
4	221 – 250	5	**16**	0.17	0.53
5	**251** – 280	**5**	21	0.17	**0.70**
6	Понад 280	9	30	0.30	1.00

1.3.2. Характеристики центру розподілу і порядкові статистики

Важливими характеристиками центру розподілу, крім середньої арифметичної, є **мода** і **медіана**, які дають змогу виявити деякі особливості структури рядів розподілу. Тому їх називають **структурними середніми. На відміну від середньої, мода і медіана не залежать від крайніх значень групувальної ознаки**, оскільки вони характеризують значення ознак, які займають певне місце в ряді розподілу.

Середня: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} = \sum_{i=1}^{n} x_i \cdot \dfrac{f_i}{\sum_{i=1}^{n} f_i} = \sum_{i=1}^{n} x_i \cdot \alpha_i,\ \sum_{i=1}^{n} \alpha_i = 1 = 100\%$.

Мода (***Мо***) – значення ознаки, що найчастіше трапляється в сукупності.

- У **дискретних рядах** мода – ознака, якій відповідною є найбільша частка.
- В **інтервальних рядах спочатку** виділяється **модальний інтервал** – інтервал, якому відповідною є найбільша частота або частка. Далі обчислюють **конкретне значення моди:**
  - в рядах із **рівними інтервалами:**

$$Mo'' = x_{Mo''} + h_{Mo''} \cdot \frac{f_{Mo''} - f_{Mo''-1}}{\left(f_{Mo''} - f_{Mo''-1}\right) + \left(f_{Mo''} - f_{Mo''+1}\right)}, \qquad (1.27)$$

де $x_{Mo''}, h_{Mo''}$ – відповідно нижня межа модального інтервалу та величина (крок) інтервалів;

$f_{Mo''-1}, f_{Mo''}, f_{Mo''+1}$ – відповідно частоти (або частки) премодального, модального та постмодального інтервалів;

– в рядах **із нерівними інтервалами** модальний інтервал визначають за найбільшою абсолютною або відносною щільністю розподілу:

$$Mo'' = x_{Mo''} + h_{Mo''} \bullet \frac{D_{в_{Mo''}} - D_{в_{Mo''-1}}}{\left(D_{в_{Mo''}} - D_{в_{Mo''-1}}\right) + \left(D_{в_{Mo''}} - D_{в_{Mo''+1}}\right)}. \quad (1.28)$$

Середина (центр) інтервалу: $x_c = \frac{x_н + x_в}{2} = x_н + \frac{h}{2}$, оскільки $x_в = x_н + h$ $0{,}5 \bullet (x_в + x_н) = 0{,}5 \bullet (x_н + x_н + h) = x_н + 0{,}5 \bullet h = x_c$.

Визначення моди в інтервальних рядах пов'язане з певною умовністю, адже значення моди в інтервальних рядах визначають лінійною інтерполяцією, виходячи з припущення про рівномірність розподілу ознаки всередині інтервалів, яке майже завжди є хибним. Отже, **з** практичного погляду **кориснішим є модальний інтервал**.

***Медіана* (*Me*)** – це значення ознаки, що ділить упорядковані одиниці сукупності за збільшенням або зменшенням значень варіаційної ознаки на дві рівні за обсягом частини із значенням ознаки, вищим і меншим за це значення.

У загальному вигляді, якщо загальна кількість одиниць сукупності n – **непарне число**, то порядковий номер одиниці сукупності в ранжованому ряді, якій належить значення медіани n_{Me}, дорівнює $n_{Me} = 0{,}5 \bullet (n+1)$.

Якщо ранжований ряд має **парну кількість одиниць**, то жодна з них не матиме значення медіани. Вона в такому разі дорівнюватиме половині суми двох значень ознак, які мають дві одиниці сукупності, що займають середнє положення у ранжованому ряду, тобто:

$$Me = \frac{x_{\frac{n}{2}} - x_{\frac{n}{2}+1}}{2}. \quad (1.29)$$

У дискретних рядах розподілу значенню медіани відповідним є значення ознаки, перша кумулятивна частота якої дорівнює чи перевищує половину обсягу сукупності, тобто $S_{fi} > 0{,}5 \bullet n$*, а для кумулятивних часток* $S_{ai} > 50\%$.

В інтервальних рядах розподілу медіану визначають у такий спосіб:

- обчислюють кумулятивні частоти або частки;
- на їхній основі, як і в дискретному ряді, визначають медіанний інтервал;
- конкретне значення медіани обчислюють за формулою

$$Me = x_{me} + h_{me} \bullet \frac{0{,}5 \bullet \sum_{i=1}^{n} f_i - S_{me-1}}{f_{me}}, \quad (1.30)$$

де x_{Me} – нижня межа медіанного інтервалу;

h_{Me} – величина медіанного інтервалу;

S_{Me-1} – кумулятивна частка або частота передмедіанного інтервалу;

f_{Me} – частота або частка медіанного інтервалу.

Математичні властивості моди та медіани

Значення моди і медіани, як і середньої арифметичної,

1) *відповідно збільшаться або зменшаться на одне й те саме число* ***A****, якщо всі варіанти збільшити або зменшити на це число* ***A***;

2) *відповідно збільшаться або зменшаться в однакову кількість разів* ***A****, якщо всі варіанти збільшити або зменшити у ту саму кількість разів* ***A***;

3) *відповідно не зміняться, якщо всі частоти або частки, помножити або поділити на будь-яке число* ***A****.*

На відміну, від середньої арифметичної, моду, *як і* ***медіану,***

4) *визначають* ***не на основі всіх значень досліджуваної ознаки, і їхня величина не залежить від крайніх значень*** *ознаки. Їхня величина зумовлюється величиною інтервалів групування;*

5) *якщо розподіл за формою наближається до нормального, то між середньої арифметичною, модою і медіаною* ***є таке наближене співвідношення:***

6) ***медіана має мінімальну властивість (центр ваги)****, яка полягає в тому, що сума абсолютних відхилень варіант від медіани – величина мінімальна, тобто менша від суми абсолютних відхилень варіант від будь-якої величини* ***A.*** Цю властивість медіани використовують під час проєктування, розміщення зупинок міського транспорту, торговельних і побутових підприємств, заготівельних пунктів, тощо, оскільки лише медіана дає можливість визначити найменшу відстань споживачів від цих об'єктів.

Для поглибленого вивчення структури рядів розподілу і виявлення її характерних особливостей важливе значення мають порядкові статистики – квантилі, які поділяють ряд розподілу на ряд рівних за чисельністю частин.

Квантилі – це значення ознаки в тих одиницях сукупності, які займають визначене порядкове місце в сукупності, упорядкованої за зростанням або зменшенням значень досліджуваної ознаки, тобто це одиниці, які подано у вигляді ранжованого ряду. За допомогою квантилів можна знайти значення ознаки, які поділятимуть ряд так:

- **чотири рівні частини** – **квартилі (*Q*)**, кожна з них містить 25% одиниць сукупності, при цьому розрізняють**:**
  - перший (нижній) квартиль Q_1 – значення ознаки в цій одиниці сукупності таке, що 25% одиниць сукупності мають значення менші, а 75% – вищі від нього;
  - медіана – середній квартиль (Q_2);

- третій (верхній) квартиль Q_3 – значення ознаки в цій одиниці сукупності таке, що 75% одиниць сукупності мають значення менші, а 25% – вищі від нього;

- **п'ять рівних частин** – **квінтилі (*q*)**, кожна з яких містить 20% одиниць сукупності, при цьому розрізняють:
  - перший квінтиль q_1 – значення ознаки в цій одиниці сукупності таке, що 20% одиниць сукупності мають значення менші, а 80% – вищі від нього;
  - другий квінтиль q_2 – значення ознаки в цій одиниці сукупності таке, що 40% одиниць сукупності мають значення менші, а 60% – вищі від нього;
  - третій квінтиль q_3 – значення ознаки в цій одиниці сукупності таке, що 60% одиниць сукупності мають значення менші, а 40% – вищі від нього;
  - четвертий квінтиль q_4 – значення ознаки в цій одиниці сукупності таке, що 80% одиниць сукупності мають значення менші, а 20% – вищі від нього;
- **10 рівних частин** – **децилі (*D*)**, тобто кожна частина містить по 10% одиниць сукупності, і децилів може бути 9. При цьому**:**
  - перший дециль D_1 – значення ознаки в цій одиниці сукупності таке, що 10% одиниць сукупності мають значення менші, а 90% – вищі від нього;
  - другий дециль D_2 – значення ознаки в цій одиниці сукупності таке, що 20% одиниць сукупності мають значення менші, а 80% – вищі від нього;
- **100 рівних частин** – **персентилі (P).**

Методика розрахунку квантилів аналогічна методиці обчислення медіани.

У дискретному ряді *розподілу квантилі – це варіант, кумулятивна частота якого дорівнює відповідному виду квантилів, квантиль може збігатись з варіантою або займати проміжне значення між двома сусідніми варіантами.*

Для інтервального ряду *розрахунок конкретного значення квантилів всередині відповідних квантильних інтервалів виконують за формулою*

$$Q_{ik} = x_{Qk} + h_{Qk} \bullet \frac{w_k \bullet \sum_{i=1}^{n} f_i - S_{fQk-1}}{f_{Qk}}, \qquad (1.31)$$

де x_{Qk} – нижня межа квантильного інтервалу;

h_{Qk} – величина квантильного інтервалу;

w_{Qk} – частка одиниць, відповідна потрібному квантилю:

($Q_1 = 0{,}25$; $q_1 = 0{,}2$; $Q_2 = 0{,}5$; $D_1 = 0{,}1$; $Q_3 = 0{,}75$; $D_8 = 0{,}8$);

$\sum_{i=1}^{n} f_i$ – сума усіх частот;

S_{fQk-1} – кумулятивна частота передквантильного інтервалу;

f_{Qk} – частота квантильного інтервалу.

Квантилі, як мода і медіана, можуть бути обчислені і за частками, і за частотами.

Методика розрахунку квантилів аналогічна методиці обчислення медіани.

***У дискретному ряді** розподілу квантилі – це варіант, кумулятивна частота якого дорівнює відповідному виду квантилів, квантиль може збігатись з варіантою або займати проміжне значення між двома сусідніми варіантами.*

***Для інтервального ряду** розрахунок конкретного значення квантилів всередині відповідних квантильних інтервалів виконують за формулою*

$$Q_{ik} = x_{Qk} + h_j \bullet \frac{w_k \bullet \sum_{i=1}^{n} f_i - S_{j-1}}{f_j} = x_{Qk} + h_j \bullet \frac{w_k \bullet n - S_{j-1}}{f_j}, \qquad (1.32)$$

де x_{Qk} – нижня межа квантильного інтервалу;

h_j – величина квантильного інтервалу;

w_k – частка одиниць, відповідна потрібному квантилю ($Q_1 = 0{,}25$; $q_1 = 0{,}2$; $Q_2 = 0{,}5$; $D_1 = 0{,}1$; $Q_3 = 0{,}75$; $D_8 = 0{,}8$);

$\sum_{i=1}^{n} f_i$ – сума усіх частот;

S_{j-1} – кумулятивна частота передквантильного інтервалу;

f_j – частота квантильного інтервалу.

> ***Приклад 1.11.*** Обчислення порядкових статистик за згрупованими даними кількості угод оренди торговельної нерухомості (див. приклад 1.10), а саме: **моди, медіани, 6-го дециля** та ($D_6 = 0{,}6$) та **18-го персентиля** та ($P_{13} = 0{,}19$). Для уточнення чисел, які слід підставити у формулу (1.32), доповнимо табл. 1.11. додатковою колонкою, за якою визначатимемо групу, у якій міститься потрібний квантиль (табл.1.13). Зазвичай для визначення порядкових статистик крайні відкриті інтервали **закривають**, **не змінюючи крок в інтервалі**.

6-й дециль вказує на межу ознаки (в розглядуваному випадку обсяг продажу валюти), яку не перевищено у 60% спостережень вибірки. З **огляду на кумулятивну частку,** 6-й дециль міститься у 5-й групі, тому

$f_j = 5$, $S_{j-1} = 16$, $x_{Qk} = 251$, $h = 30$, $n = 30$.

6-й дециль становить:

$$D = x_{Qk} + h_j \bullet \frac{w_k \bullet n - S_{j-1}}{f_j};$$

$$D = 251 + 30 \bullet \frac{0.6 \bullet 30 - 16}{5} = 251 + 12 = 263 \text{ угоди.}$$

Тобто у 60% днів вибірки кількість укладених угод оренди торговельних приміщень не перевищувала 263.

Таблиця 1.13

Вихідні дані для розрахунку порядкових статистик за формулою (1.32)

Група	Межі групи	Часто-та, f_j	Кумулятивна частота, S_{fj}	Частка, w_j	Кумулятивна частка, S_{wj}	Примітка про групу, у якій міститься шуканий квантиль
1	до 160 (131–160)	4	4	0.13	0.13	
2	161 – 190	3	7	0.10	**0.23**	*18-й персентиль*
3	191 – 220	4	11	0.13	0.37	
4	221 – 250	5	**16**	0.17	0.53	**медіанний інтервал**
5	**251** – 280	**5**	21	0.17	**0.70**	***6-й дециль***
6	понад 280 (281 – 310)	9	30	0.30	1.00	**модальний інтервал,** найбільша частота

18-й персентиль відокремлює 18% спостережень вибірки з найменшими показниками кількості взятих в оренду приміщень. З огляду на кумулятивну частку, 18-й персентиль міститься у 2-й групі, тому

f_j=3, S_{j-1}=4, x_{Qk}=161, h=30, n=30.

18-й персентиль дорівнює:

$$D = x_{Qk} + h_j \bullet \frac{w_k \bullet n - S_{j-1}}{f_j};$$

$$D = 161 + 30 \bullet \frac{0.18 \bullet 30 - 4}{5} = 161 + 8.4 \approx 170 \text{ угод.}$$

Тобто лише у 18% днів вибірки кількість укладених договорів оренди торговельних приміщень не перевищувала 170;

Медіана показує, яке значення ознаки ділить сукупність навпіл. Це 2-й квартиль, або ж 5-й дециль:

$$Me = x_{me} + h_{me} \bullet \frac{0{,}5 \bullet \sum_{i=1}^{n} f_i - S_{me-1}}{f_{me}} = 221 + 30 \bullet \frac{0{,}5 \bullet 30 - 11}{5} = 221 + 24 = 245 \text{ угод}.$$

Отже, упродовж аналізованого періоду спостерігались 15 днів, коли кількість укладених угод становила менше, ніж 245, та 15 днів, коли було укладено більшу кількість угод.

Модальний інтервал збігається з останньою 6-ю групою, подальших інтервалів немає, тому у другій різниці знаменника формули (1.28) буде підставлено 0 (f_{Mo^*+1}=0):

$$Mo'' = x_{Mo''} + h_{Mo''} \bullet \frac{f_{Mo''} - f_{Mo''-1}}{\left(f_{Mo''} - f_{Mo''-1}\right) + \left(f_{Mo''} - f_{Mo''+1}\right)};$$

$$Mo'' = 281 + 30 \bullet \frac{9 - 5}{(9 - 5) + (9 - 0)} = 281 + 9{,}2 \approx 290 \text{ угод.}$$

Таким чином, найчастіше упродовж досліджуваного 30-денного періоду трапляються дні, коли кількість укладених договорів оренди наближається до 290. Оскільки останній інтервал є найчисленнішим, розподіл ділової активності на ринку оренди торговельних приміщень не є симетричним – він **асиметричний.**

1.3.3. Характеристики варіації та форми розподілу

Варіація дає змогу встановити, як значення ознаки розміщені навколо середніх величин. На варіацію ознак впливають другорядні фактори, їхня сумісна дія і різне поєднання визначають форму та закономірність розподілу. На основі варіації оцінюють однорідність статистичної сукупності за досліджуваною ознакою, сталість індивідуальних значень цієї ознаки, типовості середньої, а також розробляють інші показники і методи вивчення соціально-економічних явищ і процесів – показники щільності зв'язку між явищами й ознаками, показники оцінок вибіркового спостереження.

Абсолютні показники варіації є іменованими величинами, тобто вони мають такі самі одиниці вимірювання, як і досліджувані ознаки.

- Найпростішим абсолютним показником варіації є **розмах варіації**, який характеризує межі, в яких змінюється значення ознаки $R=x_{max}-x_{min}$.

В **інтервальному ряду розподілу** розмах варіації визначають як різницю між верхньою межею останнього інтервалу та нижньою межею першого або як різницю між середніми цих інтервалів:

- **квартильні** ($R_Q=Q_3-Q_1$ (*50% сукупності*)) або **децильні** ($R_{D1}=D_9-D_1$ *(80% сукупності*)**,** $R_{D2}=D_8-D_2$ (*60% сукупності*)) розмахи застосовують, якщо частоти крайніх варіант ряду розподілу є малими;
- **середнє лінійне відхилення** $\overline{d}$ є середнім арифметичним з абсолютних відхилень індивідуальних значень ознаки їхньої середньої величини, причому в розрахунку беруть **модулі відхилень** (незалежно від знаків):
  - за незгрупованими даними – просте:

$$\overline{d} = \frac{\sum_{i=1}^{n} \left| x_i - \overline{x} \right|}{n};$$

  - за згрупованими даними — зважене:

$$\overline{d} = \frac{\sum_{i=1}^{n} \left| x_i - \overline{x} \right| \bullet f_i}{\sum_{i=1}^{n} f_i};$$

- **середнє квартильне відхилення** — показник варіації, який обчислюють лише на основі значень ознаки, розміщених в центральній частині ранжованого ряду:

$$\overline{Q} = \frac{\sum_{i=1}^{n} Q_3 - Q_1}{2};$$

- **дисперсія (σ^2)** становить середню арифметичну з квадратів відхилень індивідуальних значень ознаки від середньої. Чисельник дисперсії, тобто суму квадратів відхилень, називають **девіатою** і позначають як S^2. Дисперсія дає змогу усунути недолік попереднього показника, зумовлений нехтуванням знака відхилення індивідуальних значень від середнього:
  - за незгрупованими даними визначають дисперсію як просту:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n}.$$

Шляхом нескладних алгебраїчних перетворень можна отримати «**формулу сирого розрахунку»** дисперсії, яку зручно використовувати на великих вибірках :

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n} = \frac{\sum_{i=1}^{n} (x_i)^2 - 2 \cdot \overline{x} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \left(\overline{x}\right)^2}{n} = \frac{\sum_{i=1}^{n} (x_i)^2}{n} - \underbrace{2 \cdot \overline{x} \cdot \frac{\sum_{i=1}^{n} x_i}{n}}_{2 \cdot \overline{x} \cdot \overline{x}} + \frac{\sum_{i=1}^{n} \left(\overline{x}\right)^2}{n} =$$

$$= \frac{\sum_{i=1}^{n} (x_i)^2}{n} - \frac{\sum_{i=1}^{n} \left(\overline{x}\right)^2}{n} = \overline{x^2} - \overline{x}^2;$$

$$\sigma^2 = \overline{x^2} - \overline{x}^2 \qquad (1.33)$$

  - за згрупованими даними обчислюють зважену дисперсію:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 \cdot f_i}{\sum_{i=1}^{n} f_i}.$$

Навіть за умов спрощеного розрахунку дисперсії за згрупованими даними формула «сирого розрахунку» не втрачає істинності. Спрощення полягає в тому, що замість розрахунку середнього значення у кожній групі використовують значення центрів інтервалів кожної з ***k*** груп ($x_j^{ц}$). Це дає змогу знизити трудомісткість розрахунків за порівняно незначної втрати їхньої точності:

$$\sigma^2 = \frac{\sum_{j=1}^{k} \left(x_j^{ц} - \overline{x}\right)^2 \cdot f_j}{\sum_{j=1}^{k} f_j} = \underbrace{\frac{\sum_{j=1}^{k} \left(x_j^{ц}\right)^2 \cdot f_j}{\sum_{j=1}^{k} f_j}}_{\overline{x^2}} - \underbrace{2 \cdot \overline{x} \cdot \frac{\sum_{j=1}^{k} x_j^{ц} \cdot f_j}{\sum_{j=1}^{k} f_j}}_{2 \cdot \overline{x} \cdot \overline{x} = 2 \cdot \overline{x}^2} + \underbrace{\overline{x}^2 \cdot \frac{\sum_{j=1}^{k} f_j}{\sum_{j=1}^{k} f_j}}_{\overline{x}^2} \approx x^2 - \overline{x}^2 \cdot \qquad (1.34)$$

Відповідно до формул (1.33) та (1.34) дисперсію називають також **центральним моментом 2-го порядку** та позначають як μ_2**,** оскільки**:**

- **центральний момент** означає піднесення до степеня різниці між значенням ознаки й середнім значенням;
- **порядок моменту**, подібно до середнього, визначається показником степеня, до якого підносять різниці. Для дисперсії (1.33), (1.34) – це 2-й степінь.

У статистиці використовують також центральні моменти 3-го та 4-го порядків (μ_3, μ_4) – для аналізу форми розподілу, а також змішаний момент 2-го порядку (коваріацію cov(x,y)) для вивчення зв'язку між двома ознаками. Їх докладно розглянуто далі.

Математичні властивості дисперсії

1. Через зменшення всіх значень ознаки на довільну величину ***A*** дисперсія не зміниться. Це означає, що середній квадрат відхилення можна обчислювати не за вихідними значеннями ознаки, а за їхніми відхиленнями від будь-якого постійного числа ***A*** на одне й те саме число ***A***, якщо всі варіанти збільшити або зменшити на це число ***A.***

2. Якщо частоти замінити частками, то дисперсія не зміниться.

3. Якщо всі значення ознаки збільшити або зменшити в однакову кількість разів ***A***, то дисперсія відповідно збільшиться або зменшиться в кількість разів, рівну A^2:

$$\sigma^2_{(A \cdot x)} = A^2 \cdot \sigma^2_x, \ \sigma^2_{(x/A)} = \frac{\sigma^2_x}{A^2}.$$

4. Якщо обчислити середній квадрат відхилення від будь-якого постійного числа ***A***, яке відрізняється тією чи іншою мірою від середньої арифметичної ($\bar{x}$), то він завжди буде більшим за середній квадрат відхилень, обчислений від середньої арифметичної $\sigma^2_A > \sigma^2_x$**.** При цьому середній квадрат відхилень більший на певну величину – квадрат різниці між середньою і умовно взятою величиною, тобто на **(**$\bar{x} - A$**)**:

$$\sigma^2_A = \sigma^2_x + \left(\bar{x} - A\right)^2,$$

або

$$\sigma^2_x = \frac{\sum_{i=1}^{n}\left(\bar{x} - A\right)^2 \cdot f_i}{\sum_{i=1}^{n} f_i} - \left(\bar{x} - A\right)^2.$$

Отже, дисперсія від середньої завжди менша за дисперсії, обчислені від будь-яких інших величин, тобто дисперсія має властивість мінімальності.

Для того щоби отримати показник варіації, який мав би таку саму одиницю вимірювання, що й досліджувана ознака, визначають середнє

квадратичне або стандартне відхилення як корінь квадратний з дисперсії ($\sigma = \sqrt{\sigma^2}$).

Якщо розподіл варіації в сукупності наближатиметься до нормального, функцію нормального розподілу визначають інтегруванням щільності нормального розподілу:

$$F(x) = \frac{1}{\sigma_x \cdot \sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^{x} e^{-\frac{1}{2} \cdot \left(\frac{x - \bar{x}}{\sigma_x}\right)^2} dx = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^{\frac{x - \bar{x}}{\sigma_x}} e^{-\frac{t^2}{2}} dt ;$$

$$f(x) = \frac{1}{\sigma_x \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x - \bar{x}}{\sigma_x}\right)^2} .$$

Тоді можна встановити зв'язок між стандартним та середнім лінійним відхиленнями, а також стандартним відхиленням та розмахом варіації:

$$\sigma = 1{,}25 \cdot \bar{d} \Rightarrow \bar{d} = 0{,}8 \cdot \sigma,\ R = 6 \cdot \sigma .$$

Наведені залежності широко використовують в управлінні якістю продукції, зокрема правило шістьох сигм: продукція є відповідною стандарту якості, якщо мінливість її споживчих характеристик змінюється у той чи інший бік не більше, ніж на три сигми. Звичайно величина стандартного відхилення суворо регламентується нормативною документацією.

Розраховують такі коефіцієнти варіації:

- **осциляції:**

$V_R = \frac{R}{\bar{x}} \cdot 100\%$;

- **лінійний:**

$V_{\bar{d}} = \frac{\bar{d}}{\bar{x}} \cdot 100\%$;

- **квадратичний:**

$V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100\%$**,** більший за лінійний, оскільки $\sigma > \bar{d}$**.** Цей коефіцієнт використовують як критерій однорідності ознаки: ознаку вважають однорідною, якщо квадратичний коефіцієнт варіації менший за 33%, тоді середню величину оцінюють як надійну і типову;

- **квартильний:**

$V_Q = \frac{Q_3 - Q_1}{2 \cdot Me}$.

Приклад 1.12. Обчислення дисперсії, стандартного відхилення, квадратичного коефіцієнта варіації за згрупованими даними кількості угод оренди торговельної нерухомості (див. приклад 1.10). Оскільки вихідна інформація представлена у згрупованому вигляді, показник дисперсії варто обчислювати за формулою (1.34) для згрупованих даних. Завдяки цьому за досить незначної втрати

точності процедури обчислення значно прискоряться. Тому доповнимо табл. 1.11. додатковою колонкою, у якій міститься центр кожної групи, та колонками проміжних розрахунків. Зазвичай для обчислення центрів крайніх відкритих інтервалів їх закривають, не змінюючи кроків інтервалу (табл.1.14).

Середнє значення можна також обчислити за центрами груп:

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j^{ц} \cdot f_j}{\sum_{j=1}^{k} f_j} = \frac{145 \cdot 4 + 175 \cdot 3 + 205 \cdot 4 + 235 \cdot 5 + 265 \cdot 5 + 295 \cdot 9}{4 + 3 + 4 + 5 + 5 + 9};$$

$$\bar{x} = \frac{50 + 525 + 820 + 1175 + 1325 + 2655}{30} = \frac{7080}{30} = 236 \text{ угод.}$$

Отримане середнє значення дає точкову оцінку очікуваної денної кількості договорів оренди на локальному ринку торгівлі нерухомістю. Однак упродовж аналізованого 30-денного періоду не спостерігається жодного дня, в якому були укладені саме 236 угод: щоденна кількість угод у табл. 1.11 є більшою або меншою. Це свідчить про некоректність точкових оцінок та доцільність застосування оцінок інтервальних, що задають обґрунтовані межі варіації середнього значення. Способи обґрунтування інтервальних оцінок середнього викладено далі.

Таблиця 1.14

Вихідні дані та результати проміжних обчислень для розрахунку показників варіації денної кількості договорів оренди, укладених на ринку торговельної нерухомості

Група	Межі групи	Центр групи, $x_j^{ц}$	Часто-та, f_j	Відхилення від середнього $x_j^{ц} - \bar{x} = x_j^{ц} - 236$	Квадрат відхилення від середнього $(x_j^{ц} - \bar{x})^2 = (x_j^{ц} - 236)^2$	Добуток квадрата відхилення від середнього та частоти $(x_j^{ц} - \bar{x})^2 \cdot f_j = (x_j^{ц} - 236)^2 \cdot f_j$
1	до 160 (131–160)	145	4	-91	8281	33124
2	161 – 190	175	3	-61	3721	11163
3	191 – 220	205	4	-31	961	3844
4	221 – 250	235	5	-1	1	5
5	251 – 280	265	5	29	841	4205
6	понад 280 (281 – 310)	295	9	59	3481	31329
Разом			**30**			**83670**

Дисперсія становитиме:

$$D = \sigma^2 = \frac{\sum_{j=1}^{k}\left(x_j^{u} - \bar{x}\right)^2 \cdot f_j}{\sum_{j=1}^{k} f_j} = \frac{83670}{30} = 2789 \text{ угод}.$$

Розмірність дисперсії, на перший погляд, виходить за межі здорового глузду, однак лише піднесення до квадрата відхилень від середнього дає змогу коректно оцінити мінливість ділової активності на ринку торгівлі нерухомістю. Можна було б і не заповнювати останні три графи табл. 1.14, якби через це розрахункова формула не стала вкрай громіздкою:

$$\sigma^2 = \frac{\sum_{j=1}^{k}\left(x_j^{u} - \bar{x}\right)^2 \cdot f_j}{\sum_{j=1}^{k} f_j};$$

$$\sigma^2 = \frac{(145-236)^2 \cdot 4 + (175-236)^2 \cdot 3 + (205-236)^2 \cdot 4 + (235-236)^2 \cdot 5 + (265-236)^2 \cdot 5 + (275-236)^2 \cdot 9}{4+3+4+5+5+9};$$

$$\sigma^2 = \frac{(-91)^2 \cdot 4 + (-61)^2 \cdot 3 + (-31)^2 \cdot 4 + (-1)^2 \cdot 5 + 29^2 \cdot 5 + 59^2 \cdot 9}{30};$$

$$\sigma^2 = \frac{281 \cdot 4 + 3721 \cdot 3 + 961 \cdot 4 + 1 \cdot 5 + 841 \cdot 5 + 3481 \cdot 9}{30};$$

$$\sigma^2 = \frac{33124 + 11163 + 3844 + 5 + 4205 + 31329}{30} = \frac{83670}{30} = 2789.$$

Стандартне відхилення у вибірці становить:

$$\sigma = \sqrt{\sigma^2} = \sqrt{D} = \sqrt{\frac{\sum_{j=1}^{k}\left(x_j^{u} - \bar{x}\right)^2 \cdot f_j}{\sum_{j=1}^{k} f_j}} = \sqrt{\frac{83670}{30}} = \sqrt{2789} = 52{,}81 \approx 53 \text{ угоди}.$$

Квадратичний коефіцієнт варіації:

$$V_\sigma = \frac{53}{236} \cdot 100\% \approx 22{,}35 < 33\%$$

Оскільки отриманий квадратичний коефіцієнт варіації є меншим, ніж 33%, то середню кількість договорів на оренду торговельних приміщень у розмірі 236 угод на день можна вважати надійною і типовою. Крім того, за величиною квадратичного коефіцієнта варіації визначають рівень ризику економічної діяльності. Ризик вважають низьким, якщо V_σ<15%, та якщо V_σ>25%, ризик настання несприятливої події є високим. За досліджуваний період ризик стагнації ринку оренди торговельної нерухомості слід визнати як середній, оскільки 15%<22,35%<25%.

Порівнюючи значення середнього, моди та медіани, виявимо таку послідовність: 236(=$\bar{x}$)<245(=Me)<290=(Mo).

На рис. 1.3. наведено графік розподілу ділової активності на ринку торговельної нерухомості у липні–серпні 20хх року. Зафарбовані блоки відображають фактичні частоти, останній групі відповідним є найвищий блок. Загалом конфігурація стовпчастої діаграми – несиметрична: її лівий край

більш положистий та видовжений. На рис. 1.3. також наведено симетричну лінію, відповідну нормальному розподілу N($\bar{x}$ =236; σ =53). У такий скорочений спосіб записано формулу для розрахунку теоретичних частот за нормальним розподілом:

$$\hat{f}_i = \frac{e^{-\frac{1}{2}\cdot\left(\frac{x_i-\bar{x}}{\sigma}\right)^2}}{\sqrt{2\cdot\pi}} \Rightarrow \hat{f}_i = \frac{e^{-\frac{1}{2}\cdot\left(\frac{x_i-236}{53}\right)^2}}{\sqrt{2\cdot\pi}}$$

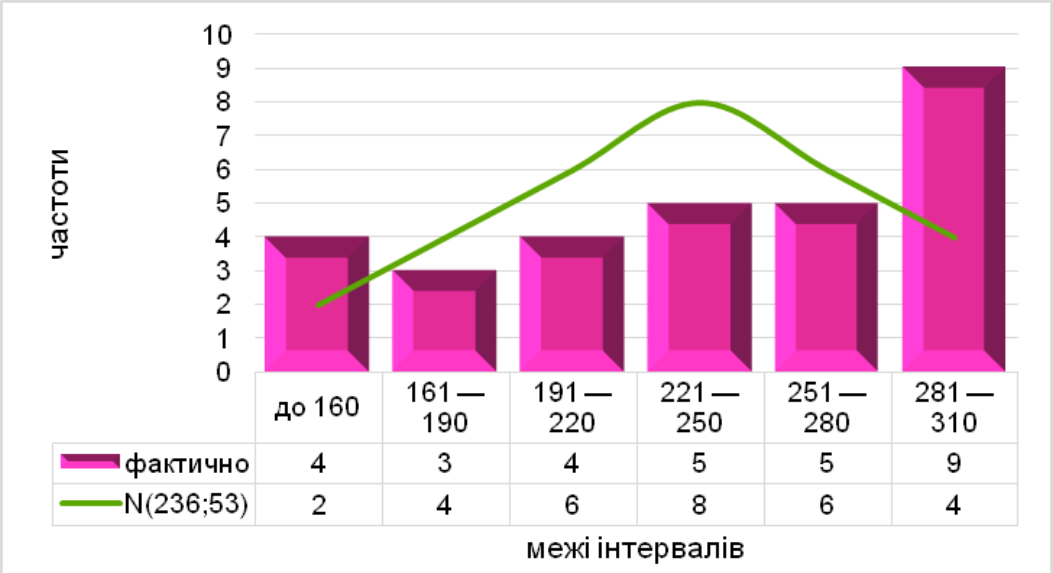


Рис. 1.3. Емпірична і теоретична криві розподілу кількості угод оренди торговельних приміщень у липні–серпні 20хх року

Для аналізу форми розподілу визначають коефіцієнт асиметрії, який може дорівнювати нулеві або бути більшим чи меншим за нього. Тобто знак цього коефіцієнта залежить від того, де будуть розміщено вищі частоти: у центру розподілу або на його краях (рис.1.4, *а* – *в*).

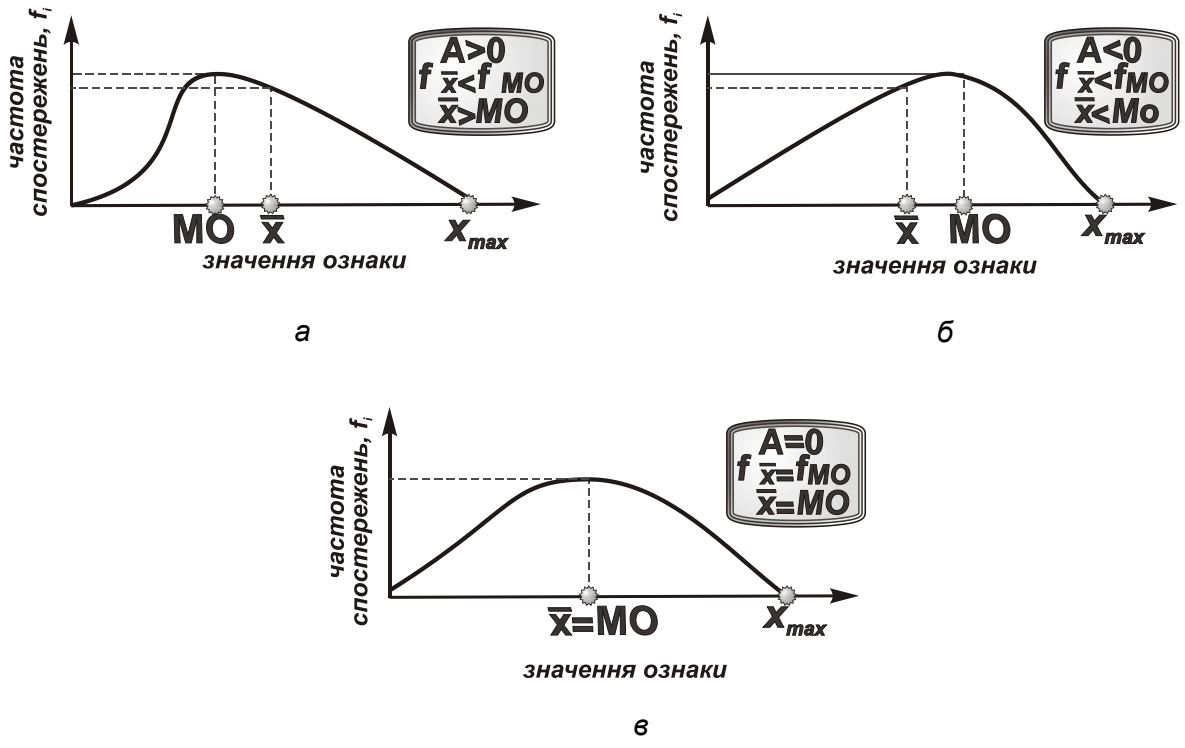


Рис. 1.4. Варіанти розподілу випадкових величин:
а – правобічна асиметрія; *б* – лівобічна асиметрія;
в – симетричний розподіл

Коефіцієнт асиметрії обчислюють за формулами:

– для незгрупованих даних

$$As = \frac{\sum_{j=1}^{k}\left(x_i - \bar{x}\right)^3}{n\sigma^3} \Rightarrow As = \frac{\mu_3}{\sigma^3}; \tag{1.35}$$

– для згрупованих даних

$$As = \frac{\sum_{j=1}^{k}\left(x_j^{u} - \bar{x}\right)^3 \cdot f_j}{\sum_{j=1}^{k} f_j \sigma^3} \Rightarrow As = \frac{\mu_3}{\sigma^3}. \tag{1.36}$$

Дріб у чисельниках формул (1.35), (1.36) – центральний момент 3-го порядку для згрупованих та незгрупованих даних. Оскільки різниці підносяться до третього степеня, момент μ_3 отже, і коефіцієнт асиметрії може набувати як додатних, так і від'ємних значень:

– якщо більші частоти розміщені для інтервалів, де ознака вища за середню, результати формул (1.35), (1.36) – від'ємні:

As<0, асиметрія лівобічна.

Лівобічна асиметрія характеризується видовженою лівою гілкою та стрімкою правою (рис. 1.4, б), В такому разі значення моди й модальної частоти є більшими відповідно за середню та частоту, якій середня належить:

$$\bar{x} < \text{Mo};$$

– навпаки, якщо більші частоти розміщені для інтервалів, у яких ознака нижча за середню, результати формул (1.35), (1.36) – додатні:

As>0, асиметрія правобічна.

Правобічна асиметрія характеризується стрімкою лівою гілкою та видовженою правою гілкою (рис. 1.4, *а*) В цьому випадку значення моди й модальної частоти є меншими за середню та частоту, якій середня належить:

$$\text{Mo} < \bar{x};$$

– коли більші частоти розташовані в тому самому інтервалі, що й середня, результати формул (1.35), (1.36) близькі до нуля:

$\text{As} \approx 0$, розподіл симетричний.

У такому разі графік емпіричних частот симетричний (рис.1.4, *в*), модальна частота одночасно є й частотою, що містить середню, а значення самої середньої та моди збігатимуться:

$$\text{Mo} \approx \bar{x}.$$

Таким чином, під час визначення асиметрії на основі графіків частот слід бути особливо уважним, оскільки тип асиметрії протилежний розміщенню максимуму. Візуальний аналіз рис. 1.3. свідчить, що емпіричний розподіл ділової активності на ринку оренди торговельної нерухомості – асиметричний лівобічний. Загальна конфігурація рис. 1.3. подібна до рис. 1.4, *б*.

Іншою характеристикою форми розподілу є ексцес, який змінюється залежно від гостроти вершини графіка емпіричного чи теоретичного розподілу. Нормальний розподіл має ексцес, рівний 3:

$$\text{Es}_{N(\bar{x},\sigma)} = 3. \qquad (1.37)$$

У статистичній літературі його обґрунтовують, інтегруючи моменти 2-го та 4-го порядку щільності нормального розподілу. За рівномірного розподілу, коли всі частоти рівні, а крива розподілу являє собою горизонтальну лінію, ексцес є меншим за 3. Це – плосковершинний розподіл. У статистичних дослідженнях та математичному моделюванні широко використовують розподіл Стьюдента. Він симетричний, але має гострішу вершину. Відповідно його ексцес перевищує 3.

Для емпіричного розподілу ексцес розраховують за такими формулами:

– для незгрупованих даних:

$$Es = \frac{\frac{\sum_{j=1}^{k}\left(x_i - \bar{x}\right)^4}{n}}{\sigma^4} \Rightarrow Es = \frac{\mu_4}{\sigma^4}; \quad (1.38)$$

– для згрупованих даних:

$$Es = \frac{\frac{\sum_{j=1}^{k}\left(x_j^u - \bar{x}\right)^4 \cdot f_j}{\sum_{j=1}^{k} f_j}}{\sigma^4} \Rightarrow Es = \frac{\mu_4}{\sigma^4}. \quad (1.39)$$

Дріб у чисельниках формул (1.38), (1.39) – центральний момент 4-го порядку (μ_4) для згрупованих та незгрупованих даних. Він, як і дисперсія, завжди додатний, оскільки різниці підносяться до четвертого степеня.

Розподіл гостровершинний, якщо результат обчислень за виразами (1.38) чи (1.39) перевищує 3:

Es>3⇒ розподіл гостровершинний.

Та коли ексцес (1.38), (1.39) є меншим, ніж 3, розподіл вважають плосковершинним:

Es>3⇒ розподіл плосковершинний.

На основі асиметрії та ексцесу можна дуже наближено встановити, чи є близьким до нормального емпіричний розподіл. З цією метою розраховують **стандартне відхилення ексцесу й асиметрії** за формулами, аргументом яких є лише обсяг вибірки:

– стандартне відхилення асиметрії:

$$\sigma_{As} = \sqrt{\frac{6 \cdot (n-2)}{(n+1) \cdot (n+3)}}; \quad (1.40)$$

– стандартне відхилення ексцесу:

$$\sigma_{Es} = \sqrt{\frac{24 \cdot n \cdot (n-2) \cdot (n-3)}{(n+1)^2 \cdot (n+3) \cdot (n+5)}}. \quad (1.41)$$

Після цього визначають критерії значущості асиметрії та ексцесу шляхом ділення модулів асиметрії та ексцесу на відповідні стандартні відхилення, обчислені за виразами (1.40), (1.41).

Розподіл є відповідним нормальному, якщо отримані співвідношення не перевищують 3:

$$\frac{|As|}{\sigma_{As}} < 3, \frac{|Es|}{\sigma_{Es}} < 3$$

$$\Downarrow$$

розподіл близький до нормального

$$\Downarrow$$

асиметрія та ексцес неістотні

Подібно до правила шістьох сигм в управлінні якістю, якщо наведені співвідношення перевищать 3, варто вважати емпіричний розподіл далеким від нормального:

$$\frac{|As|}{\sigma_{As}} > 3, \frac{|Es|}{\sigma_{Es}} > 3$$

$$\Downarrow$$

розподіл далекий від нормального

$$\Downarrow$$

асиметрія та ексцес істотні

Наведений підхід до перевірки відповідності емпіричного розподілу нормальному має досить наближений, умовний характер, проте він демонструє процедуру верифікації статистичних гіпотез. Такі процедури, як і приклад перевірки відповідності емпіричного закону нормальному, наведено в другому розділі.

Приклад 1.13. Обчислення показників форми розподілу за згрупованими даними про кількість угод оренди торговельної нерухомості (див. приклад 1.10). Використовуючи дані табл. 1.14, складемо таблицю допоміжних розрахунків (табл. 1.15).

Коефіцієнт асиметрії (1.36) становить:

$$As = \frac{\frac{\sum_{j=1}^{k}\left(x_j^u - \bar{x}\right)^3 \cdot f_j}{\sum_{j=1}^{k} f_j}}{\sigma^3} = \frac{\frac{-1625970}{3}}{\left(\sqrt{2789}\right)^3} \Rightarrow As = \frac{\mu_3}{\sigma^3} =$$

$$= \frac{-54199}{147289.83} = -0{,}368 < 0 \Rightarrow \text{лівобічна.}$$

Оскільки коефіцієнт асиметрії — від'ємний, асиметрія **лівобічна**, що підтверджує рис. 1.3.

Таблиця 1.15

Вихідні дані та результати проміжних обчислень для визначення показників форми розподілу денної кількості договорів оренди, укладених на ринку торговельної нерухомості

Група	Межі групи	Центр групи, $x_j^ц$	Частота, f_j	Відхилення від середнього $x_j^ц - \bar{x} = x_j^ц - 236$	Добуток кубу відхилення від середнього та частоти $(x_j^ц - \bar{x})^3 \cdot f_j = (x_j^ц - 236)^3 \cdot f_j$	Добуток четвертого ступеню відхилення від середнього та частоти $(x_j^ц - \bar{x})^4 \cdot f_j = (x_j^ц - 236)^4 \cdot f_j$
1	до 160 (131–160)	145	4	-91	–1507142	137149922
2	161 – 190	175	3	-61	–907924	55383364
3	191 – 220	205	4	-31	–178746	5541126
4	221 – 250	235	5	-1	–8	8
5	251 – 280	265	5	29	146334	4243686
6	Понад 280 (281 – 310)	295	9	59	821516	48469444
Разом			**30**		**–1625970**	**250787550**

Ексцес (1.39) становить:

$$Es = \frac{\frac{\sum_{j=1}^{k}(x_j^ц - \bar{x})^4 \cdot f_j}{\sum_{j=1}^{k} f_j}}{\sigma^4} = \frac{\frac{250787550}{30}}{2789^2} \Rightarrow Es = \frac{\mu_4}{\sigma^4} =$$

$$= \frac{8359585}{7778521} = 1{,}075 < 3 \Rightarrow \text{плосковерш инний.}$$

Здобутий результат ексцесу є значно меншим за 3, тому аналізований розподіл слід визнати **плосковершинним.**

Під час обчислення асиметрії та ексцесу враховано такі елементарні залежності:

$$\sigma^3 = \left(\sigma^2\right)^{\frac{3}{2}} = \left(\sqrt{\sigma^2}\right) = \left(\sqrt{2789^2}\right)^3; \quad \sigma^4 = \left(\sigma^2\right)^2 = 2789^2.$$

Приклад 1.14. Встановлення міри наближеності форми розподілу нормальному закону кількості угод оренди торговельної нерухомості за даними й результатами прикладів 1.10 – 1.13.

Стандартні відхилення ексцесу й асиметрії обчислюють за формулами (1.40), (1.41), n=30, оскільки наявні 30 спостережень:

– стандартне відхилення асиметрії:

$$\sigma_{As} = \sqrt{\frac{6 \cdot (n-2)}{(n+1) \cdot (n+3)}} = \sqrt{\frac{6 \cdot (30-2)}{(30+1) \cdot (30+3)}} = \sqrt{\frac{6 \cdot 28}{31 \cdot 33}} = 0{,}405;$$

– стандартне відхилення ексцесу:

$$\sigma_{Es} = \sqrt{\frac{24 \cdot n \cdot (n-2) \cdot (n-3)}{(n+1)^2 \cdot (n+3) \cdot (n+5)}} = \sqrt{\frac{24 \cdot 30 \cdot (30-2) \cdot (30-3)}{(30+1)^2 \cdot (30+3) \cdot (30+5)}} = \sqrt{\frac{24 \cdot 30 \cdot 28 \cdot 27}{31^2 \cdot 33 \cdot 35}} = 0{,}70 \cdot$$

На підставі знайдених стандартних відхилень обчислюють критеріальні співвідношення:

– для асиметрії:

$$\frac{|As|}{\sigma_{As}} = \frac{|-0{,}368|}{0{,}405} = 0{,}909 < 3 \Rightarrow \text{розподіл наближений до нормального}$$

$$\Downarrow$$

асиметрія неістотна;

– для ексцесу:

$$\frac{|Es|}{\sigma_{As}} = \frac{|1{,}075|}{0{,}700} = 1{,}536 < 3 \Rightarrow \text{розподіл наближений до нормального}$$

$$\Downarrow$$

ексцес неістотний.

Результати розрахунків радикально суперечать результатами побудови графіка емпіричних частот (див. рис.1.3), адже стовпчаста діаграма не відтворює дзвоноподібної конфігурації теоретичної лінії нормального розподілу: висота стовпчиків у більшості груп нижча за теоретичну лінію, а останній стовпчик є зависоким.

Результати прикладу 1.14. підтверджують умовний характер перевірки наближеності розподілу до нормального на основі показників ексцесу й асиметрії.

1.4. ІНТЕРВАЛЬНІ ОЦІНКИ ХАРАКТЕРИСТИК РЯДІВ РОЗПОДІЛУ. ВИБІРКОВЕ СПОСТЕРЕЖЕННЯ

1.4.1. Інтервальна оцінка середніх значень у великих вибірках

Дані про розподіл значень ознаки у сукупності можуть бути отримані в разі дослідження не всієї генеральної сукупності, а тільки її частини – **вибірки**. Обсяг вибірки визначають на основі статистичного аналізу. Загалом вибіркове обстеження дає досліднику можливість визначити середню арифметичну вибіркової сукупності $\overline{x}$, а також величину граничної похибки цієї середньої Δx, яка з певною імовірністю показує, наскільки вибіркова середня може відрізнятися від середньої генеральної сукупності $\overline{x}$ (всієї генеральної сукупності опитаних) у більший чи менший бік. Верхня границя генеральної

середньої $\bar{x} + \Delta\bar{x}$, нижня – $\bar{x} - \Delta\bar{x}$, проте можна об'єднати обидві границі в один запис:

$$\bar{x} \pm \Delta\bar{x}. \tag{1.42}$$

Надійний (довірчий) інтервал – межі, в яких з певною імовірністю може бути невідома величина оцінюваного параметра. Він має ще одну назву – гранична помилка середнього. Гранична помилка середнього для великих вибірок, обсяг яких є більшим за 30 одиниць (**n>30**) визначають так:

$$\Delta\bar{x} = Z^* \cdot \frac{\sigma}{\sqrt{n}}, \tag{1.43}$$

де Z* – коефіцієнт довіри, визначений зі статистичних таблиць кумулятивної функції нормального розподілу Ф(Z*). Алгоритм обчислення цієї функції залежно від значень аргументу – імовірності помилкового визначення надійного інтервалу – також реалізовано в Excel і подібних табличних процесорах. Тому альтернативним способом визначення коефіцієнта довіри є використання функції табличного процесора Excel НОРМСТОБР, або ж NORM.S.INV в англомовних версіях.

Значок «зірочка» біля літерного позначення вказує на те, що така величина може бути визначена із статистичних таблиць. Її називають також **квантилем розподілу**, або **табличним** значенням**,** або **критичним** значенням. Табульовано не лише нормальний розподіл, а й деякі інші, використовувані в статистичному аналізі. Тому зірочку вживають і біля інших буквених символів, що позначають певний закон розподілу.

Безпомилково визначити межі надійного інтервалу середнього значення неможливо за будь-яких умов, оскільки розрахунки виконують на основі окремої вибірки даних, а не цілої сукупності. Немає жодної гарантії, що в деякій іншій вибірці, утвореній зовсім іншими спостереженнями, буде таке значення середнього показника, яке вийде за межі надійного інтервалу, обґрунтованого на іншій вибірці. Тому, визначаючи надійний інтервал, дослідник самостійно визначає також імовірність помилки середнього. Зазвичай в економічних, зокрема маркетингових, дослідженнях імовірність помилки беруть на рівні $\alpha = 0{,}05$. Це відповідне середній точності дослідження: лише у 5 зі 100 подібних вибірок значення середнього буде вищим або нижчим, ніж $\bar{x} \pm \Delta\bar{x}$. Для досліджень в техніці, біології, медицині звичайно потрібна вища точність, тому ймовірність помилки обмежують $\alpha = 0{,}01$ (висока точність) й $\alpha = 0{,}001$ (дуже висока точність). Якщо ймовірність помилки становить $\alpha = 0{,}1$**,** це свідчить про низьку точність. У дослідженнях економічних процесів теж можуть підвищуватись вимоги до точності.

Оскільки за виразом (1.43) визначають півширину надійного інтервалу, тобто модуль відхилення, що може як підвищувати, так і зменшувати

середнє, результатом функції кумулятивного нормального розподілу (Φ(Z*)) є різниця між 1 та половиною імовірності помилки:

$$\Phi(Z^*) = 1 - \frac{\alpha}{2}. \qquad (1.44)$$

Тобто імовірність помилки порівну ділиться між недо- та переоцінкою середнього значення.

Якщо визначати **Z*** за допомогою функції табличного процесора Excel NORM.S.INV, її аргументом під час обґрунтування півширини надійного інтервалу також слід задавати величину $1 - \frac{\alpha}{2}$, набираючи у клітинці

$$= \text{NORM.S.INV}\left(1 - \frac{\alpha}{2}\right).$$

Другий співмножник формули (1.43), що містить дріб, у знаменнику якого квадратний корінь з кількості спостережень, має також назву стандартне відхилення середнього.

Стандартне відхилення середнього, яке позначають як $\sigma_{\bar{x}}$, розраховують за формулою

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\sigma^2}}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}. \qquad (1.45)$$

Для того щоби уникнути помилок округлення, у вираз (1.45) доцільно підставляти показник дисперсії. При цьому дріб під квадратним коренем, що має назву «дисперсія середнього», позначають як $\sigma^2_{\bar{x}}$ або $D_{\bar{x}}$.

На основі формули граничної помилки (1.43) у наукових та маркетингових дослідженнях розв'язують три типи обернених задач.

1-й тип задач

Знайти півширину надійного інтервалу (граничну помилку) за відомими даними про значення ознаки у деякій вибірці обсягом ***n***, на основі яких можна оцінити дисперсію ознаки (σ^2). Крім того, сам дослідник визначає межі істинності інтервальної оцінки середнього для генеральної сукупності, за якими на основі статистичних таблиці визначають ***Z****. Для розв'язання такої задачі, власне, й використовують формулу (1.43).

> ***Приклад 1.15.*** Обґрунтування інтервальної оцінки середньоденної кількості угод оренди торговельної нерухомості за вихідними даними й результатами прикладів 1.10 – 1.14.

Якщо взяти до уваги висновки прикладу 1.14 про несуттєвість асиметрії та ексцесу, можна вважати, що ділова активність на ринку оренди нерухомості є відповідною нормальному закону. Досліджувана вибірка охоплювала 30 днів (n=30). Вибіркова статистика:

- точкова оцінка середньої кількості угод $\overline{x}$ =236;
- дисперсія: $\sigma^2 = D = 2789$;

Побудуємо два варіанти імовірності помилки, виходячи із різних вимог до надійності результату

- імовірність помилки становитиме $\alpha_1 = 0{,}05$, відповідний їй коефіцієнт довіри – *Z_1*=1,96*. Його отримано за допомогою функції Excel: $= \text{NORM.S.INV}\left(1 — 0{,}05/2\right)$. Такий самий результат можна було б отримати, відшукавши в таблиці кумулятивної функції стандартного нормального розподілу число 0,975 та встановивши десяті й соті знаки *Z** як відповідні назві рядка й стовпчика. Подібну таблицю представлено у наступному розділі посібника;
- буде взята також інша імовірність помилки $\alpha_2 = 0{,}001$, відповідний коефіцієнт довіри якої *Z_2**=2,58. З цією метою у функції Excel задано інший аргумент: $= \text{NORM.S.INV}\left(1 — 0{,}01/2\right)$.

За формулою (1.43) півширина надійного інтервалу становитиме:

– для імовірності помилки $\alpha_1 = 0{,}05$:

$$\Delta\overline{x_1} = Z_1 * \bullet \frac{\sigma}{\sqrt{n}} = 1{,}96 \bullet \frac{\sqrt{2789}}{\sqrt{30}} = 1{,}96 \bullet 9{,}642 \approx 19 \text{ угод};$$

– для імовірності помилки $\alpha_2 = 0{,}01$:

$$\Delta\overline{x_2} = Z_2 * \bullet \frac{\sigma}{\sqrt{n}} = 2{,}58 \bullet \frac{\sqrt{2789}}{\sqrt{30}} = 2{,}58 \bullet 9{,}642 \approx 25 \text{ угод}.$$

Інтервальна оцінка середньоденної кількості орендних угод за формулою (1.42) становить:

– для імовірності помилки $\alpha_1 = 0{,}05$:

$$\overline{x} \pm \Delta\overline{x}_1 = 236 \pm 19 \text{ угод}.$$

Це означає, що у 95 зі 100 інших 30-денних вибірок спостережень за діловою активністю на ринку комерційної нерухомості середньоденна кількість укладених договорів може бути інакшою, ніж 236, але не меншою за 217 й не більшою за 255. Лише п’ять вибірок із 100 характеризуватимуться середньоденною кількістю укладених орендних угод, меншою за 217 або ж більшою за 255;

– для імовірності помилки $\alpha_2 = 0{,}01$:

$$\overline{x} \pm \Delta\overline{x}_2 = 236 \pm 25 \text{ угод}.$$

У 99 зі 100 інших 30-денних вибірок спостережень за укладанням орендних угод на ринку комерційної нерухомості середньоденна кількість укладених договорів перебуватиме в межах від 211 до 261. І тільки в одному випадку зі 100 середня кількість угод виявиться меншою, ніж 211, чи більшою, аніж 261.

Зниження імовірності помилки з 0,05 до 0,01 призвело до розширення меж надійного інтервалу. Тому замалі значення імовірності помилки (довірчої імовірності) можуть призвести до надмірно широких надійних інтервалів, за яких оцінка середнього стане малоінформативною. Особливо значною втрата інформаційної цінності буде у тому випадку, коли надійний інтервал міститиме нуль, тобто ліва й права межі матимуть різні знаки.

2-й тип задач

Знайти **імовірність помилки**, якщо відомі півширина надійного інтервалу ($\Delta\bar{x}$), дисперсія у вибірці (σ^2), обсяг вибірки (***n***). Запитання таких задач зазвичай формулюють так: «З якою імовірністю можна стверджувати, що середнє перебуває в межах від… до …».

Алгоритм розв'язання задач цього типу

1. За відомим вірогідним інтервалом та стандартним відхиленням середнього розраховують квантиль нормального розподілу:

$$Z^* = \frac{\Delta\bar{x}}{\sigma_{\bar{x}}}. \qquad (1.46)$$

2. На основі квантиля **(Z*)** нормального розподілу визначають функцію стандартного нормального розподілу: **Ф(Z*).** Для цього використовують таблиці стандартного нормального розподілу або функцію Excel НОРМСТРАСП з групи СТАТИСТИЧНІ. Аргументом цієї функції є число-квантиль нормального розподілу, здобуте на етапі 1. В останніх англомовних версіях табличного процесора цій функції відповідною є інша, дещо модифікована **NORM.S.DIST**. У ній, крім квантиля, є додатковий параметр **«сукупне»**, який визначає, чи потрібно визначити кумулятивну функцію, якщо задано 1, чи щільність розподілу, якщо користувач ввів 0. Таким чином, слід в клітинці процесора набрати: **=NORM.S.DIST(Z з етапу 1;1).**

3. Імовірність помилки (α) розраховують за формулою

$$\alpha = 2 \bullet (1 - \Phi(Z^*)). \qquad (1.47)$$

Оскільки півширина надійного інтервалу середнього поширюється на два боки (до більших і менших значень) формула (1.47) містить співмножник «2».

Таким чином, імовірність можна визначити зі статистичних таблиць. Для цього треба знати Z*. Оскільки стандартне відхилення – корінь квадратний з дисперсії (позначуваної як σ^2 або D), вираз (1.45) можна замінити еквівалентними залежностями:

$$Z^* = \frac{\Delta\bar{x} \bullet \sqrt{n}}{\sigma} = \frac{\Delta\bar{x} \bullet \sqrt{n}}{\sqrt{\sigma^2}} = \frac{\Delta\bar{x} \bullet \sqrt{n}}{\sqrt{D}}. \qquad (1.48)$$

Приклад 1.16. З якою імовірністю можна стверджувати, що середня ціна 1-кімнатної квартири перебуває в межах від 1,55 млн грн до 2,23 млн грн, якщо **дисперсія** вибірки – 722 500 (тис. грн)^2 (D=722500), обсяг вибірки – 36 об'єктів (n=36).

Межі середнього – подвійна півширина надійного інтервалу – 0,68 млн грн (=2,23–1,55), або 680 тис. грн, адже розмірність дисперсії пов'язана із тисячами гривень, а не з мільйонами.

Півширина надійного інтервалу: $\Delta\bar{x}$ =340 тис. грн (=680/2).

Застосовуємо поданий алгоритм:

1. Квантиль нормального розподілу (1.48):

$$Z^* = \frac{\Delta\bar{x} \bullet \sqrt{n}}{\sqrt{D}} = \frac{340 \bullet \sqrt{36}}{\sqrt{722500}} = 2{,}4 .$$

2. Функцію стандартного нормального розподілу:

$$\Phi(Z^*)=\Phi(2{,}4)=0{,}9918.$$

Це значення отримано за допомогою табличного процесора Excel шляхом введення в одну з клітинок формули

=NORM.S.DIST(2,4;1).

3. Імовірність помилки (α)за формулою (1.46) становить:

$$\alpha = 2 \bullet (1 - \Phi(Z^*)) = 2 \bullet (1 - 0{,}9918) = 2 \bullet 0{,}0082 = 0{,}0164 .$$

Отже, обсяг вибірки та її дисперсія дають підстави стверджувати що середня ціна 1-кімнатної квартири обчислюється в межах від 1,55 млн грн до 2,23 млн грн з імовірністю $\alpha = 0{,}0164$. Ціна буде в таких межах для 984 об'єктів з 1000 (оскільки $1 - \alpha = 0{,}9836$).

3-й тип задач

Знайти обсяг вибірки, достатньої для заданої імовірності помилки, якщо відомі півширина надійного інтервалу та дисперсія у вибірці.

Алгоритм розв'язання задач цього типу

1. Оскільки імовірність помилки відома (задається дослідником), то з таблиць знаходять Z* відповідно до умови (1.46), тобто: $\Phi(Z^*)=1-0{,}5 \bullet \alpha$. Альтернативою таблицям є функції табличного процесора Excel **НОРМСТОБР** або **NORM.S.INV**.

2. Далі обсяг вибірки обчислюють за формулою, отриманою на підставі виразу (1.43):

$$n = \left(\frac{Z^* \bullet \sigma}{\Delta \bar{x}}\right)^2 = \left(\frac{Z^*}{\Delta \bar{x}}\right)^2 \bullet \sigma^2 = \left(\frac{Z^*}{\Delta \bar{x}}\right)^2 \bullet D, \qquad (1.49)$$

адже слід визначити невідомий дільник:

$$\Delta \bar{x} = Z^* \bullet \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{Z^* \bullet \sigma}{\Delta \bar{x}}\right)^2 .$$

Приклад 1.17. Яким має бути обсяг вибірки, щоби на рівні імовірності помилки $\alpha = 0{,}05$ гранична помилка ціни квартир становила 12 тис. грн ($\Delta \bar{x} = 12$). Дисперсія ціни квартир згідно з попередніми дослідженнями становить *800 (тис. грн)^2 (D=800).*

1. Відповідно до формули (1.46) знаходимо табличне значення кумулятивної функції стандартного нормального розподілу та її квантиль:

$$\Phi(Z^*) = 1 — 0{,}5 \bullet 0{,}05 = 1 — 0{,}025 = 0{,}975;$$

$$Z^* = 1{,}96 \ (=NORM.S.INV(1\text{-}0.05/2)).$$

2. Обсяг вибірки згідно з формулою (1.49):

$$n = \left(\frac{Z^*}{\Delta \bar{x}}\right)^2 \bullet D = \left(\frac{1{,}96}{12}\right)^2 \bullet 800 = 21{,}34 \approx 22 \text{ оголошення} .$$

Зазвичай розв'язки таких задач заокруглюють до більшого цілого числа, адже це підвищує статистичну значущість розрахунків, про що йтиметься далі.

Отриманий результат – 22 оголошення не відображає проблеми **повторності**, коли одне й те саме оголошення буде враховане кілька разів. Для усунення цієї проблеми розглядають безповторні вибірки, при цьому формула для обчислення обсягу таких вибірок буде дещо складнішою, ніж вираз (1.49).

1.4.2. Методи формування вибіркових сукупностей

Інтервальна оцінка середнього втрачає свою інформативну цінність у міру розширення меж інтервалу. Тимчасом межі інтервалу «розсуваються» внаслідок збільшення граничної помилки середнього. Аналізуючи склад формули для її обрахунку (1.43), можна виявити три фактори зменшення півширини надійного інтервалу

$$\Delta \bar{x} \downarrow = \frac{Z^* \downarrow \bullet \sigma \downarrow}{\sqrt{n \uparrow}} :$$

- **зменшення коефіцієнта довіри** ($Z^* \downarrow \Rightarrow \Delta \bar{x} \downarrow$, це скоротить величину в чисельнику й результат співвідношення). Втім, цей спосіб є неприйнятним, оскільки зростає імовірність помилки ($Z^* \downarrow \Rightarrow \Phi(Z^*) \downarrow \Rightarrow \alpha \uparrow = 2 \bullet (1 — \Phi(Z^*) \downarrow)$, отже, знижується точність оцінки або, у термінах статистики, її значущість;

- **зменшення дисперсії** й стандартного відхилення у вибірці ($\sigma \downarrow \Rightarrow \Delta\bar{x} \downarrow$, оскільки зменшення чисельника зменшить весь дріб). Однак здійснити це майже неможливо, оскільки у більшості досліджень дисперсія сукупності, як і її середні, є невідомою. Натомість доводиться орієнтуватись на розрахунки дисперсії у вибірках пробних досліджень;
- **збільшення обсягу вибірки (**$n \uparrow \Rightarrow \Delta\bar{x} \downarrow$, оскільки дріб зменшується у разі зростання знаменника**).** Це – єдиний спосіб одночасного підвищення інформаційної цінності за збереження статистичної значущості інтервальної оцінки. Крім того, внаслідок збільшення обсягу вибірки оцінка дисперсії у сукупності виявиться точнішою, оскільки будуть охоплені більше варіант прояву досліджуваної ознаки.

Проте збільшення обсягу вибірки потребує більших затрат на виконання дослідження: більше часу доведеться витратити на обстеження об'єктів, спостереження, а в маркетингових дослідженнях – на анкетування респондентів. Фінансовий аспект збільшення обсягу вибірки – додаткові витрати коштів на підготовку й організацію спостережень. Насамперед зростуть витрати на друк анкет, забезпечення тривалішого доступу до сайтів електронних форм опитування, оплату праці інтерв'юерів і працівників, залучених до оброблення результатів анкетування.

Ефективність статистичного дослідження й, відповідно, витрат на його виконання суттєво знижується, якщо досліджувана вибірка буде з повторами. При цьому спотворюються дані про реальний розподіл ознак у сукупності. Особливо яскраво це проявляється в маркетингових дослідженнях, коли один й той самий респондент на ті самі питання може як свідомо, так і несвідомо давати різні відповіді.

Для обґрунтування **обсягу безповторної вибірки** використовують дещо видозмінену формулу граничної помилки середнього. Видозміна полягає в тому, що результати вивчення ознаки у вибірці обсягом **n** поширюють на решту сукупності, тобто на *N–n* спостережень, позначаючи великою буквою *N* чисельність генеральної сукупності. Тоді відносну частку генеральної сукупності, не охоплену дослідженнями, визначають так:

$$\frac{N - n}{N} = 1 - \frac{n}{N}.$$

При цьому для безповторної вибірки, коли один об'єкт не може бути досліджений декілька разів, формула граничної помилки середнього така:

$$\Delta\bar{x} = Z^* \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} = Z^* \cdot \sqrt{\frac{D}{n}\left(1 - \frac{n}{N}\right)}. \qquad (1.50)$$

На практиці часто задають співвідношення вибірки й сукупності, якщо вибірка 10%-ва: $\frac{n}{N} = 0{,}1$.

Та коли співвідношення вибірки й сукупності є невідомим чи не заданим, обсяг вибірки визначають на основі перетворення формули (1.50):

$$\Delta\bar{x}^2 = (Z^*)^2 \bullet \frac{D}{n}\left(1 - \frac{n}{N}\right) \Rightarrow N \bullet \Delta\bar{x}^2 = \frac{(Z^* \bullet \sigma)^2}{n} \bullet (N - n);$$

$$\Downarrow$$

$$n = \frac{N \bullet (Z^* \bullet \sigma)^2}{(Z^* \bullet \sigma)^2 + N \bullet \Delta\bar{x}^2}. \qquad (1.51)$$

Оскільки стандартне відхилення визначається на основі дисперсії, то для уникнення додаткових помилок внаслідок заокруглення квадратних коренів у вираз (1.52) доцільно підставити значення дисперсії (*D*):

$$n = \frac{N \bullet D \bullet (Z^*)^2}{D \bullet (Z^*)^2 + N \bullet \Delta\bar{x}^2}. \qquad (1.52)$$

Застосування формул (1.51), (1.52) потребує таких даних:

1. Дисперсія у сукупності ($\sigma^2 \equiv D$) чи стандартне відхилення (σ) – ці показники встановлюють за даними попередніх досліджень або на підставі припущень.

2. Імовірність помилки (α) дослідник задає залежно від бажаної точності оцінок. Найчастіше α набуває значень 0,05, 0,01, 0,001. Імовірність помилки може бути виражена також у відсотках: 5%, 1%, 0,1%.

3. Межі помилки середнього ($\bar{x} \pm \Delta\bar{x}$), або півширина надійного інтервалу, ($\Delta\bar{x}$) – визначає дослідник як гіпотетичні, підтвердити чи спростувати які й має статистичний експеримент.

4. Обсяг генеральної сукупності (*N*). В економічних дослідженнях генеральною сукупністю може бути чисельність населення певного регіону, міста, району, або кількість розташованих там нерухомих об'єктів чи об'єктів рухомого майна, весь обсяг випуску продукції, кількість господарських операцій. Використовуючи формулу (1.51), треба підставляти *повний запис N.* Тобто відповідне число слід записувати з усіма нулями, інакше підстановка кількості тисяч чи мільйонів призведе до хибних «скорочень» чисельника й знаменника та суттєво спотворить результат, занижуючи потрібний обсяг вибірки. Генеральною сукупністю може бути також деякий проміжок часу, тоді вибіркою буде деякий фрагмент цього проміжку і саме формула (1.51) і визначатиме величину цього фрагменту.

Приклад 1.18. Яким має бути обсяг вибірки, щоби на рівні імовірності помилки $\alpha = 0{,}001$ гранична помилка ціни квартир становила п'ять тисяч гривень ($\Delta\bar{x} = 5$). Стандартна помилка ціни квартир, згідно з попередніми дослідженнями, 20 000 грн

($\sigma = 20$ тис. грн **,** одиниця виміру збігається з вимірником досліджуваної ознаки та її граничної помилки, тобто у розрахунках використовують тисячі гривень). На інтернет-порталі агентства нерухомості є в наявності близько 23 тисяч оголошень з пропозицією подібної нерухомості (***N=23000***). **Увага! *N* підставляють повністю, без скорочень, тобто з усіма нулями.**

Для того щоби визначити обсяг вибірки за формулою (1.51), потрібно встановити квантиль нормального розподілу ***Z****. Для цього слід врахувати співвідношення (1.46) та скористатись таблицею кумулятивної функції стандартного нормального розподілу або функцією Excel **NORM.S.INV:**

Z*=3,29 *(=NORM.S.INV(1-0.001/2)).*

Згідно з формулою (1.51) обсяг вибірки становить:

$$n = \frac{N \bullet (Z^* \bullet \sigma)^2}{(Z^* \bullet \sigma)^2 + N \bullet \Delta\bar{x}^2} = \frac{23000 \bullet (3{,}29 \bullet 20)^2}{23000 \bullet 5^2 + (3{,}29 \bullet 20)^2} = \frac{99581720}{579329.6} \approx 172 \text{ оголошення}. \quad (1.53)$$

Підстановка до виразу (1.51) **23 тис.** замість **N=23000** могла б призвести до помилки, оскільки чисельник виразу (1.53) скоротився б у 1000 разів, у той час як у знаменнику скорочення стосувалося б лише першого доданка. Таке некоректне використання вхідних даних дало б багаторазове зменшення обсягу вибірки проти правильного результату розв'язання рівняння (1.53).

Приклад 1.19. Визначити тривалість періоду роботи ринку оренди торговельної нерухомості, який слід проаналізувати з погляду стабільності попиту серед орендарів, якщо у розпорядженні є архівні дані за останні 10 років, щороку – 250 робочих днів. Дослідження має підтвердити думку про те, що в середньому за день укладається від 230 ($\bar{x} - \Delta\bar{x}$) до 240 ($\bar{x} + \Delta\bar{x}$) договорів оренди. Допустима (гранична) імовірність помилки $\alpha = 0{,}05$. Попередні пробні дослідження (див. приклад 1.12) показали, що дисперсія середньоденної кількості угод становить $\sigma^2 = D = 2789$**.**

У цьому прикладі обсяг вибірки – кількість днів, щодо яких слід дослідити кількість укладених орендних угод для подальшого уточнення інтервальної оцінки середнього. Оскільки в умові наведено показник дисперсії, застосовуємо вираз (1.52). Тобто за формулою (1.52) буде визначена кількість робочих днів, адже вибіркою є частина 10-річного часового проміжку. Цей 10-річний період і є генеральною сукупністю, вимірюваною робочими днями. Тому обсяг генеральної сукупності становить:

N=10 років•250 робочих днів/рік=2500 робочих днів.

Відповідно до умови задачі ширина надійного інтервалу – різниця між верхньою та нижньою межею кількості укладених договорів: ***10 угод (=240–230)***.

Тоді півширина інтервалу, тобто гранична помилка: $\Delta\bar{x} = 5$ угод $\left(= \frac{10}{2}\right)$.

Для застосування формули (1.52) визначаємо коефіцієнт довіри *Z**, використовуючи при цьому співвідношення (1.46) й статистичні таблиці кумулятивної функції стандартного нормального розподілу або їхню альтернативу – функцію Excel **NORM.S.INV:**

Z*=1,96 *(=NORM.S.INV*(1-0.05/2)).

Обсяг вибірки, таким чином, становить:

$$n = \frac{N \bullet D \bullet (Z^*)^2}{D \bullet (Z^*)^2 + N \bullet \Delta\bar{x}^2} = \frac{2500 \bullet 2789 \bullet 1{,}96^2}{2789 \bullet 1{,}96^2 + 2500 \bullet 5} \approx 1154 \text{ дні}.$$

Для прискорення розрахунків замість значення *Z**=1,96 правомірним буде підставити 2. Звісно, при цьому дещо зросте й обсяг вибірки, проте таке незначне зростання кількості досліджуваних об'єктів позитивно позначиться на точності результатів, оскільки квантилеві нормального розподілу *Z**=2 відповідною буде імовірність помилки не 0,05, а дещо менша (1.46):

$$\alpha = 2 \bullet (1 - \Phi(Z^*)) = 2 \bullet (1 - NORM.S.DIST(2;1)) = 2 \bullet 0{,}002275 = 0{,}0455 .$$

Тобто одночасно із прискоренням обчислювальних процедур зростає і надійність результатів: імовірність помилки за *Z**=2 становитиме не 5%, а тільки 4,55%.

Для прикладу 1.19 обсяг вибірки в такому разі буде:

$$n = \frac{N \bullet D \bullet (Z^*)^2}{D \bullet (Z^*)^2 + N \bullet \Delta\bar{x}^2} = \frac{2500 \bullet 2789 \bullet 2^2}{2789 \bullet 2^2 + 2500 \bullet 5} = 1179 \approx 1180 \text{ днів}.$$

Отже, для досягнення 95%-ї точності оцінки середнього рівня ділової активності ринку оренди торговельної нерухомості, коли помилка оцінки середньоденної кількості угод не перевищить п'яти, слід проаналізувати кількість угод, укладених упродовж чотирьох років і дев'ятьох місяців. Звичайно, кращі результати будуть за умови, коли 1180 днів становитимуть не суцільний період, а будуть випадковим чином розподілені по всіх 10 роках. Для цього доцільно випадковим чином обрати перший день, включений до вибірки (наприклад, згенерувавши випадкове число від 1 до 2500, а потім вирахувати відповідний день з усього періоду, умовно пронумерувавши дні). Оскільки обсяг вибірки наближається до половини сукупності, то кожен наступний елемент вибірки віддалятиметься від попереднього на два дні. Та якщо буде досягнутий останній день сукупності, відбір слід вести далі від початку 10-річного періоду. В цьому полягає простий механічний відбір елементів вибірки.

Якщо визначають обсяг вибірки для меж **частки,** формула має такий вигляд:

$$n = \frac{N \bullet (Z^*)^2 \bullet w \bullet (1 - w)}{(Z^*)^2 \bullet w \bullet (1 - w) + N \bullet (\Delta w)^2}, \qquad (1.53)$$

де w – частка об'єктів сукупності, що мають певну ознаку. У статистичних дослідженнях ця частка може бути визначена за припущеннями дослідника або за результатами попередніх експериментів. Дослідження якраз і полягає в обґрунтуванні меж такої частки;

w•(1–w) – дисперсія частки;

Δw – гранична помилка частки, тобто півширина надійного інтервалу частки, значення якої дослідник обирає самостійно, зважаючи на інформативність оцінки.

Відмінність виразу (1.53) від (1.51) полягає у заміні дисперсії вибірки (σ^2**)** на дисперсію частки (w•(1–w)).

Приклад 1.20. Для маркетингового дослідження споживання полівінілацетатних фарб в м. Суми потрібно провести анкетування з надійністю 99,7% .

Обчислити обсяг вибірки, якщо відомо, що населення Сум станом на 1 березня 2015 року становило 268 409 осіб (*N*), з яких 35% (*w*=0,35) не купують цієї продукції. При цьому, за даними попередніх досліджень, похибка частки коливається в межах 2,8% ($\Delta w = 0{,}028$).

- **Коефіцієнт довіри** (квантиль нормального розподілу): ***Z*=2,97***. Результат визначено за допомогою введення в табличну клітинку Excel такої функції:

***=NORM.S.INV*(1-(1-0,997)/2).**

В аргументі функції табличного процесора враховано формулу (1.46), застосування якої ґрунтується на тому, що за надійності дослідження 99,7% ($\gamma = 0{,}997$) імовірність помилки становитиме:

$$\alpha = 1 - \gamma = 1 - 0{,}997 = 0{,}003 ;$$

- обсяг вибірки визначають за виразом (1.53):

$$n = \frac{N \cdot (Z^*)^2 \cdot w \cdot (1 - w)}{N \cdot \Delta w^2 + (Z^*)^2 \cdot w \cdot (1 - w)} = \frac{268409 \cdot 2{,}97^2 \cdot 0.35 \cdot (1 - 0.35)}{268409 \cdot 0.028^2 + 2{,}97^2 \cdot 0.35 \cdot (1 - 0.35)}$$

$$n = \frac{538631.036}{212.44} \approx 2536 \text{ покупців.}$$

Подібно до прикладу 1.19 значення коефіцієнта довіри доцільно заокруглити, вважаючи, що *Z**=3. Оскільки внаслідок заокруглення квантиль збільшується, імовірність помилки зменшується, отже, зростає надійність дослідження. При цьому дещо збільшиться й обсяг вибірки:

$$n = \frac{N \cdot (Z^*)^2 \cdot w \cdot (1 - w)}{N \cdot \Delta w^2 + (Z^*)^2 \cdot w \cdot (1 - w)} = \frac{268409 \cdot 3^2 \cdot 0.35 \cdot (1 - 0.35)}{268409 \cdot 0.028^2 + 3^2 \cdot 0.35 \cdot (1 - 0.35)}$$

$$n = \frac{549567.4}{212.48} = 2586.441 \approx 2587 \text{ покупців.}$$

Якщо жодних припущень про значення частки *w* зробити неможливо, попередні дослідження не були виконані, значення частки вважають рівним

0,5 (*w*=0,5***)***, що є відповідним максимальній мірі невизначеності, коли відношення шансів «50 на 50».

> ***Приклад 1.21.*** Про скількох покупців акрилових фарб слід мати інформацію, щоби встановити питому вагу задоволених якістю продукції із точністю до 1,5% і надійністю 0,954, не маючи жодної інформації про враження від продукції. Загальна кількість покупців продукції за останній рік у всіх точках продажу – 48,6 тис. осіб.

Вихідні дані для формули (1.53)*: N=48600 осіб,* $\Delta w = 0{,}015$.

Оскільки немає жодних підстав для визначення частки задоволених продукцією покупців, беруть ***w=0,5***.

Відповідно до наведених прикладів ***Z*=2***, результат, отриманий за допомогою табличного процесора шляхом введення у клітинку функції **=NORM.S.INV(1-(1-0.954)/2)**, заокруглено.

В аргументі функції NORM.S.INV застосовано формулу (1.46), причому за надійності дослідження 95,4% ($\gamma = 0{,}954$) імовірність помилки становитиме:

$$\alpha = 1 - \gamma = 1 - 0{,}954 = 0{,}045 .$$

Обсяг вибірки за формулою (1.53) дорівнює:

$$n = \frac{N \cdot (Z^*)^2 \cdot w \cdot (1 - w)}{N \cdot \Delta w^2 + (Z^*)^2 \cdot w \cdot (1 - w)} = \frac{972000 \cdot 2^2 \cdot 0.5 \cdot (1 - 0.5)}{972000 \cdot 0.015^2 + 2^2 \cdot 0.5 \cdot (1 - 0.5)};$$

$$n = \frac{972000}{219.7} = 4424.3 \approx 4425 \text{ покупців.}$$

1.4.3. ІНТЕРВАЛЬНА ОЦІНКА СЕРЕДНІХ ЗНАЧЕНЬ У МАЛИХ ВИБІРКАХ

Малою вважають вибірку, обсяг якої не перевищує 30 спостережень: $n \le 30$. Однак залежно від змісту статистичної гіпотези вимоги до розміру вибірки можуть суттєво змінюватись. Наприклад, для визначення інтервальної оцінки частки вибірка має бути більшою за 60 об'єктів, а обсяг спостережень, менший за 60, визнається як мала вибірка, непридатна для оцінювання частки в сукупності. Інтервальне оцінювання частки виконують за формулою, аналогічною виразу (1.43), використовуючи замість стандартного відхилення вибірки (σ)стандартне відхилення частки ($\sqrt{w \cdot (1 - w)}$).

Для визначення інтервальної оцінки середнього на малих вибірках слід мати на увазі дві принципові обставини, які унеможливлюють застосування формули (1.43):

1. Припущення про розподіл ознаки за нормальним законом є справедливим лише для великих вибірок. На основі вивчення малої вибірки стверджувати про нормальність розподілу помилок середнього неможливо.

Помилка середнього – це різниця між значенням варіанти та середнім, і для малих вибірок їхній розподіл вважають відмінним від нормального, тому використання ***Z**** як коефіцієнта довіри є некоректним. З цією метою використовують інший розподіл – розподіл Стьюдента (t-розподіл), а коефіцієнт довіри позначають як ***t****.

2. Через невеликий обсяг спостережень ускладнюється оцінювання варіації ознаки в сукупності. Тому у **малих вибірках потрібно забезпечити незміщуваність оцінки дисперсії**, тобто уникнути її заниження. При цьому беруть до уваги показник кількості ступенів вільності середнього (n—1), за допомогою якого визначають виправлену дисперсію:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}. \qquad (1.54)$$

Стандартне відхилення ознаки у малих вибірках також визначають як виправлене:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}. \qquad (1.55)$$

Беручи до уваги вираз (1.54) і раніше наведені формули, можна встановити зв'язок між дисперсією сукупності та виправленою дисперсією малої вибірки. Іншими словами, для переходу від невиправленої дисперсії до виправленої слід виконати таке коригування:

$$s^2 = \frac{\sigma^2}{n-1} \bullet n \Rightarrow s^2 = \sigma^2 \bullet \frac{n}{n-1}. \qquad (1.56)$$

Аналогічне виразу (1.56) коригування виконують і для «виправлення» стандартного відхилення, але при цьому потрібно добути квадратний корінь:

$$s = \sigma \bullet \sqrt{\frac{n}{n-1}}. \qquad (1.57)$$

Подібні коригування слід виконати і для «виправлення» результату формули «сирого розрахунку»:

$$s^2 = \left(\overline{x^2} - \bar{x}^2\right) \bullet \frac{n}{n-1}. \qquad (1.58)$$

Тобто прискорити розрахунки стандартного відхилення малої вибірки можна у такий спосіб:

$$s = \sqrt{\left(\overline{x^2} - \bar{x}^2\right) \bullet \frac{n}{n-1}}. \qquad (1.59)$$

Через згадані обставини півширину надійного інтервалу середнього для малих вибірок визначають за іншою формулою:

$$\Delta\bar{x} = t^* \bullet \frac{s}{\sqrt{n}}, \qquad (1.60)$$

де t^* – квантиль розподілу Стьюдента, який визначають за спеціальними статистичними таблицями цього розподілу. Приклад такої таблиці та її

використання наведено в наступному розділі. В останніх версіях табличних процесорів для виразу (1.54) слід використовувати формулу T.INV.2T, яка має два аргументи: не лише ймовірність помилки (α), а й кількість ступенів вільності $\nu = n - 1$. Тобто до клітинки табличного процесора вводять вираз типу = T.INV.2T **(**α; $n - 1$**),** підставляючи замість букв потрібні числа. Сутність поняття «кількість ступенів вільності» проілюстровано далі на конкретному прикладі;

s – стандартна помилка малої вибірки, яку отримують, коригуючи стандартну помилку на кількість ступенів вільності, за формулою (1.57).

Взагалі, виправляти дисперсію й стандартне відхилення відповідно до кількості ступенів вільності слід для будь-яких вибірок. Однак для великих вибірок коригувальні множники дисперсії $\left(\frac{n}{n-1}\right)$ та стандартного відхилення $\left(\sqrt{\frac{n}{n-1}}\right)$ наближають до 1, тому ними нехтують.

Зокрема, якщо *n* = 1:

– коригувальний множник дисперсії становить 1,034$\left(=\frac{30}{30-1}\right)$;

– для стандартного відхилення – лише 1,017$\left(=\sqrt{\frac{30}{30-1}}\right)$.

Тимчасом як для вибірок, що містять менш ніж 10 одиниць, значення обох коригувальних множників є значно вищими за 1, а тому використання невиправленої дисперсії спричинить помилкове зниження надійного інтервалу.

Під поняттям «кількість ступенів вільності» в статистиці розуміють кількість варіант значення ознаки, які можуть довільно, тобто необмежено, змінюватись. Зокрема, у середнього значення кількість ступенів вільності є на 1 меншою за обсяг вибірки, тому для інтервальної оцінки середнього й використовують величину $\nu\ (\nu = n - 1)$. Значення «кількості ступенів вільності» середнього можна пояснити на такому прикладі.

> ***Приклад 1.22.*** Зроблено припущення, що середня трудомісткість робіт із мурування 1 м^3 цегляної кладки становить 7 люд.-год. Для його перевірки заплановано виконати п’ять вимірювань. Виконано лише чотири вимірювання, результати такі:
> 1-ше випробування: 8,0 люд.-год (x1=8,0);
> 2-ге випробування: 6,8 люд.-год (x2=6,8);
> 3-тє випробування: 6,4 люд.-год (x3=6,4);
> 4-те випробування: 7,2 люд.-год (x4=7,2).

Коли буде завершено 5-те випробування, одержимо x_5, причому воно буде вирішальним для підтвердження чи спростування припущення про 7 люд.-год середньої трудомісткості:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{8{,}0 + 6{,}8 + 6{,}4 + 7{,}2 + x_5}{5} = 7{,}0. \qquad (1.61)$$

Рівність (1.61) буде справедливою лише за одного значення x_5: $x_5 = 7{,}0 \cdot 5 - (8{,}0 + 6{,}8 + 6{,}4 + 7{,}2) = 6{,}6$ люд. – год.

Якщо x_5 буде іншим, середня трудомісткість також матиме інше значення, а не 7 люд.-год, як припускалось. Таким чином, довільно змінюватись можуть тільки чотири варіанти вибірки, а 5-та суворо детермінується їх значеннями, тобто визначається алгебраїчним рівнянням. Отже, кількість ступенів у вибірці цього прикладу – чотири, на один менша за обсяг вибірки: $\nu = n - 1 = 5 - 1 = 4$.

Приклад 1.23. Виконати інтервальну оцінку стійкості діамантових дисків для нарізування затверділого залізобетону на рівні значущості $\alpha = 0{,}01$. Під час виконання робіт був повністю використаний ресурс дев'ятьох дисків. У приблизно однакових умовах роботи тривалість їхнього використання до повного спрацювання становила:

1-й диск – 140 год. (**x_1=140**); 2-й диск – 130 год. (**x_2=130**);
3-й диск – 135 год. (**x_3=135**); 4-й диск – 145 год. (**x_4=145**);
5-й диск – 155 год. (**x_5=155**); 6-й диск – 110 год. (**x_6=120**);
7-й диск – 115 год. (**x_7=115**); 8-й диск – 160 год. (**x_8=160**);
9-й диск – 125 год. (**x_9=125**).

Точкова оцінка середньої стійкості алмазного інструмента:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{140 + 130 + 135 + 145 + 155 + 110 + 115 + 160 + 125}{9} = 135 \text{ год.}$$

Виправлене стандартне відхилення (дисперсію визначено методом «сирого розрахунку») згідно з формулою (1.59):

$$s = \sqrt{\left(\overline{x^2} - \bar{x}^2\right) \cdot \frac{n}{n-1}};$$

$$s = \sqrt{\left(\frac{140^2 + 130^2 + 135^2 + 145^2 + 155^2 + 110^2 + 115^2 + 160^2 + 125^2}{9} - 135^2\right) \cdot \frac{9}{9-1}};$$

$$s = \sqrt{287{,}5} = 16{,}96 \approx 17 \text{ год.}$$

Невиправлена стандартна помилка була б меншою в 1,06 раза $\left(1{,}06 = \sqrt{1{,}125}\right)$, це б звузило надійний інтервал середнього, а разом з ним і статистичну значущість інтервальної оцінки.

Коефіцієнт довіри: $t^*_{(\alpha;\nu=n-1)} = t^*_{(0{,}01;9-1)} = t^*_{(0{,}01;8)} = 3{,}355$.

Табличне значення *t**-критерію визначено за допомогою Excel, до клітини якого введено формулу: = T.INV.2T(0,01;9-1**)**. Результат виявився більшим за 3, у той час, як Z* =2,58<3 за імовірності $\alpha = 0{,}01$. В цьому й полягає одна з відмінностей малої вибірки від великої: неможливість застосування закону нормального розподілу помилок й заміна його розподілом Стьюдента сприяє збільшенню коефіцієнта довіри.

Півширина надійного інтервалу за формулою (1.60):

$$\Delta\bar{x} = t^* \bullet \frac{s}{\sqrt{n}} = 3{,}355 \bullet \frac{17{,}0}{\sqrt{9}} \approx 19 \text{ год.}.$$

Інтервальна оцінка середньої стійкості алмазного інструмента для нарізання затверділого залізобетону на рівні значущості $\alpha = 0{,}01$ становить 135 ± 19 годин.

Отже, у 99 інших подібних вибірках зі 100 середня стійкість алмазного диска перебуватиме в межах від 116 до 154 год безвідмовної роботи.

ЗАВДАННЯ ДЛЯ САМОСТІЙНОГО ОПРАЦЮВАННЯ МАТЕРІАЛУ

1. За даними прикладу 1.10.

- 1-й і 3-й квартилі, 1-й і 4-й квітнилі, 1-й і 9-й децилі, 68-й персентиль. Дайте пояснення результатам.
- Обчисліть середню з групових середніх для зазначеного ряду за формулами середньої гармонійної і середньої квадратичної.

2. Визначте обсяг виконаних будівельних робіт в умовно-натуральних показниках, якщо відомо, що за квартал виконано робіт з простого фарбування стін 11,3 тис. м^2, поліпшеного – 16,9 тис. м^2, високоякісного – 29,7 тис. м^2. Затрати праці на 100 м^2 фарбування підготовлених поверхонь становлять для простих стін – 100 люд.-год, для поліпшених – 50 люд.-год, для високоякісних – 80 люд.-год.
3. Обчисліть середньорічний індекс обсягу благ у домогосподарствах за період від 2004 до 2016 р. (табл. 1.16). Скористайтися формулою середнього геометричного.

Таблиця 1.16

Наявність у домогосподарствах окремих товарів тривалого користування (у середньому на 100 домогосподарств, штук)

Благо	2004	2006	2008	2010	2012	2014	2016
мобільні телефони	15	81	149	167	187	197	201

4. Обчисліть середньорічні темпи зростання вартості поїздки в метро, якщо відомо, що в 2014 році вартість поїздки становила 2 грн, а в 2018 –

   8 грн. Поясніть результати.
5. Згрупуйте ряд даних про забезпеченість населення транспортними засобами і на підставі згрупованих даних обчисліть середню з групових середніх, моду, всі квартилі, 85-й персентиль.
6. Для маркетингового дослідження попиту на мармелад з груш в м. Полтава потрібно провести анкетування, за результатами якого гранична помилка середнього річного споживання цього солодкого продукту не перевищувала б 300 г з надійністю 95,4%. Обчислити обсяг вибірки, якщо відомо, що населення Полтави станом на 1 березня 2016 року становило 293 тис. осіб, а відповідно до попереднього дослідження стандартне відхилення річного обсягу споживання грушевого мармеладу становить 6,5 кг.
7. Використовуючи дані завдання № 5 (табл. 1.17), дайте інтервальну оцінку забезпеченості населення транспортними засобами:
   - на рівні значущості $\alpha = 0{,}05$ у країнах, назва яких починається з букви «П»;

- на рівні значущості $\alpha = 0{,}025$ у країнах, назва яких починається з букви «І».

Увага! Визначаючи числові показники для завдання № 5, слід взяти:

- h* – остання цифра номера вашого паспорта;
- g* – передостання цифра номера вашого паспорта.

Якщо біля числового показника немає буквених доданків чи співмножників, такий показник є спільний для всіх варіантів.

Таблиця 1.17

Забезпеченість населення колісними транспортними засобами (крім мотоциклів) у 20хх р. (шт. на 1000 населення)

Країна	Кіль-кість авто	Країна	Кіль-кість авто	Країна	Кіль-кість авто	Країна	Кіль-кість авто
Ісландія	866	Фінляндія	795	Австралія	752	Японія	718
Камбоджа	208	Іспанія	710	Люксембург	787	Канада	667
Південна Африка	213	Чехія	605	Норвегія	754	Італія	854
Данія	535	Словенія	666	Австрія	754	Угорщина	376
Великобританія	577	Ірландія	546	Франція	664	Мексика	318
Бельгія	637	Греція	866	Німеччина	687	Південна Корея	459
Нідерланди	600	Польща	664	Швеція	618	Ямайка	131
США	875	Литва	506	Марокко	102	Аргентина	552
Нова Зеландія	765	Португалія	549	Малайзія	828	Сербія	292
Швейцарія	734	Ізраїль	361	Чилі	251	Нігерія	84

8. Для дослідження річних витрат домогосподарств на ремонт житла виконано (10+h*)% вибіркове опитування в різних областях центрального економічного регіону України. В результаті отримано такий розподіл домогосподарств за витратами на ремонт житла:

Середня сума річних витрат на ремонт житла, тис. грн	до 5	5–8,9	9–12,9	понад 13
Кількість респондентів	30–g*	20+h*	40–g*	5

Для безповторної вибірки визначте:

– з імовірністю 0,954 граничну похибку вибіркової середньої і межі середньої суми витрат на ремонт житла за генеральною сукупністю;

– з імовірністю 0,997 граничну похибку вибірки для визначення частки і межі питомої ваги домогосподарств, витрати на ремонт яких перевищують 4 тис. грн;

– обсяг вибіркової сукупності за умови, що гранична похибка частки домогосподарств із річними витратами на ремонт понад 8 000 грн з імовірністю 0,954 не перевищувала 5%;

– обсяг вибіркової сукупності за умови, що гранична похибка вибірки у визначенні середньої суми витрат на ремонт житла з імовірністю 0,997 не перевищувала 300 грн.

Знайдіть моду, медіану, четвертий квінтиль.

Знайдіть третій квартиль, дев’ятий дециль, (70—g*—h*)-й персентиль.

Розділ 2
МЕТОДИ ПЕРЕВІРКИ СТАТИСТИЧНИХ ГІПОТЕЗ

2.1. ІНСТРУМЕНТАРІЙ ТА ПРОЦЕДУРА ПЕРЕВІРКИ СТАТИСТИЧНИХ ГІПОТЕЗ

Жодне ґрунтовне дослідження в економіці не може бути виконане без висунення статистичних гіпотез. Дослідники розрізняють наукові та статистичні гіпотези.

Статистична гіпотеза – це будь-яке висловлювання про характер змін випадкової величини у генеральній сукупності, яку перевіряють за результатами випадкових спостережень. Процедуру зіставлення висловленої гіпотези з даними вибірки називають **перевіркою статистичної гіпотези**. Ця процедура дає відповідь на запитання, **чи є спостережувані результати випадковими, чи вони реальні**, що дає змогу на основі наявної інформації зробити вибір між двома гіпотезами (припущеннями). Про перевірку статистичних гіпотез іноді говорять як про спосіб використання статистики для ухвалення рішень.

Методи перевірки статистичних гіпотез ґрунтуються на перевірці певних підстав, що належать до генеральної сукупності (популяції). Якщо ці підстави можуть стосуватися числових характеристик сукупностей, то ми маємо справу з параметричною верифікацією. У разі припущення щодо функціонального розподілу верифікація має непараметричний характер.

Наприклад, якщо гіпотезу сформульовано так:

- «нормальний закон має задану дисперсію, або середню величину» – це параметрична гіпотеза;
- «варіаційний ряд розподілу має нормальний закон розподілу» – це непараметрична гіпотеза.

Основні види висловлених під час статистичного опрацювання даних гіпотез стосуються такого:

- однорідність двох чи декількох опрацьовуваних вибірок або деяких характеристик аналізованих сукупностей;
- тип закону розподілу досліджуваної випадкової величини;
- стаціонарність і незалежність опрацьовуваної низки спостережень;
- тип залежності між складовими досліджуваної багатомірної ознаки.

Нульова гіпотеза **(нуль-гіпотеза, H_0)** – статистична гіпотеза, яку перевіряють. Параметрична нульова гіпотеза стосується параметрів розкладів (або чисел, що характеризують популяцію), записують її як H_0: параметр, що дорівнює закладеному числу (x=160 грн, $\sigma = 50$ грн – середній чек дорівнює 160 грн зі стандартним відхиленням 50 грн).

Гіпотеза, що їй суперечить – **альтернативна (конкурентна, H_1)**. Вона може мати три варіанти:

- H_1: параметр не дорівнює закладеному числу – двобічна гіпотеза (середній чек не дорівнює 160 грн і може бути як більшим, так і меншим);
- H_1: параметр більший від закладеного числа. – однобічна (правобічна) гіпотеза (середній чек перевищує 160 грн);
- H_1: параметр менший від закладеного числа. – однобічна (лівобічна) гіпотеза (середній чек є меншим за 160 грн).

Непараметричні гіпотези

Нульова гіпотеза (нуль-гіпотеза, *H_0*)

H_0:* популяція має розклад *G (продуктивність праці мулярів розподілена за нормальним законом, інтенсивність поломок фарбопульта відповідна закону Пуассона). *Впливу заходів з підвищення енергоефективності на витрати немає, зміна собівартості продукції виникла внаслідок випадкових факторів.*

- Гіпотеза, що їй суперечить, – **альтернативна (конкурентна, *H_1)* найчастіше двобічна.**

H_1:* популяція не має розкладу *G (продуктивність праці мулярів розподілена не за нормальним законом, інтенсивність поломок фарбопульта не узгоджується із законом Пуассона *(розподіл кожної із зазначених ознак підпорядковується будь-якому іншому закону розподілу, аніж припущено: рівномірному, експоненційному, логнормальному)**)**. Внаслідок заходів з підвищення енергоефективності відбувається суттєва зміна собівартості продукції (проте не уточнюється, в якому напрямі відбувається така зміна: ощадливість, чи перевитрати). Якби уточнювалось, що відбувається – **економія (лівобічна),** чи **перевитрати (правобічна)** – **тоді альтернативні гіпотези були б однобічними***.

Ставлячи гіпотезу H_0, можна ухвалити одне з двох рішень: прийняти H_0, або спростувати H_0. Одночасно така гіпотеза може мати дві логічні властивості: справжня H_0, або фальшива H_0 (табл. 2.1).

Таблиця 2.1

Основні типи помилок у верифікації статистичних гіпотез

Характеристика гіпотези	**Нуль-гіпотеза, *H_0*, істинна**	**Нуль-гіпотеза, *H_0*, хибна**
Нуль-гіпотеза, *H_0*, **приймається**	Помилки немає, **висока імовірність уникнення помилки**	**Помилка другого виду, імовірність β (1— β)– потужність**
Нуль-гіпотеза, *H_0*, **відхиляється**	**Помилка першого виду, імовірність α – значущість**	Помилки немає **висока імовірність уникнення помилки**

- **Помилка першого виду** – нульову гіпотезу відкидають, коли вона насправді правильна. Імовірність припуститися помилки першого виду позначають як α**. Це число** α називають **рівнем значущості.** Висновок формулюють так: «Нульову гіпотезу відкинуто на рівні значущості α». **Найчастіше** α **= 0,1; 0,05; 0,02; 0,01; 0,001** залежно від рівня точності дослідження та ризику втрат через ухвалення помилкової гіпотези *(переоцінений попит на товар і завеликі залишки нереалізованих запасів за своїми наслідками незрівнянні із запуском у виробництво лікарського препарату із недооціненим побічним ефектом).* Не придбали потрібної книги; відмовили претенденту на роботу, хоча його кандидатура відповідала всім вимогам; вирішили, що партія товару бракована, хоча неякісним був лише один виріб (пара взуття).
- **Помилка другого виду** – нульову гіпотезу приймають, коли насправді вона хибна. Імовірність помилки другого виду позначають як β**. Потужність критерію** – величина **1**$-\beta$ **дорівнює** імовірності відкинути неправильну гіпотезу. Придбали непотрібну книгу; взяли на роботу претендента, який не відповідає вимогам до компетенції; вирішили, що партія товару задовільної якості, коли насправді вся вона бракована (тобто вирішили, що неякісним був лише один виріб (пара взуття), який потрапив до рук).

Перевірку статистичних гіпотез виконують за допомогою різноманітних **статистичних критеріїв** – деяких випадкових величин, щодо яких відомі закони розподілу і які можуть бути обчислені на підставі наявних даних. Найпоширеніші статистичні критерії розроблено на основі таких відомих розподілів:

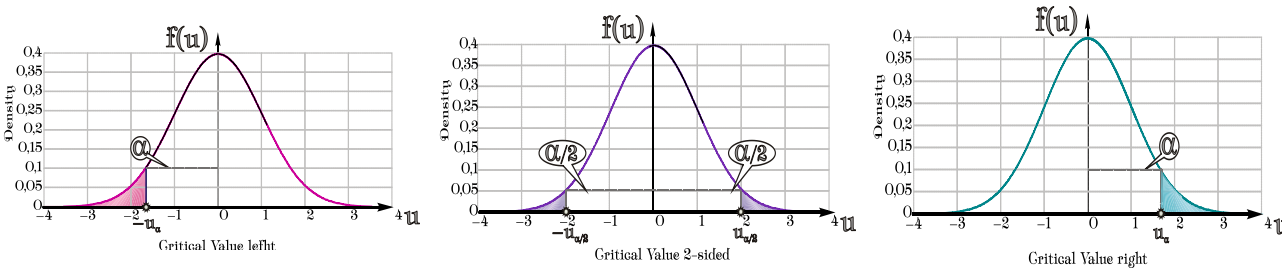
- χ^2 – хі-квадрат;
- Стьюдента;
- Фішера.

Для цих критеріїв складено таблиці, у яких зазначено критичні точки, відповідні певному рівню значущості та кількості ступенів свободи.

Серед безлічі можливих значень критерію обирають підмножину – критичну область. ***Критична область*** – інтервал у якому найчастіше трапляється переважна більшість значень критерію. ***Критичні точки*** – це граничні точки критичної області. Критичні точки встановлюють відповідно до визначеного рівня значущості α та максимального рівня потужності (1$-\beta$). Тобто критична область є відповідною абсцисам (значенням критерію), для яких величина імовірності перевищує рівень істотності (значущості) α (який визначають залежно від цілей дослідження).

Можливими є три види розміщення критичної області (залежно від виду нульової й альтернативної гіпотез, виду і розподілу статистичного критерію), що представлено на рис. 2.1.

- **Двобічна критична область** – якщо альтернативну гіпотезу сформульовано у такому вигляді:
  H_1: параметр не дорівнює закладеному числу. Тоді критична область складається з двох інтервалів: $\left(-\infty; x^{кр}_{лів.\frac{\alpha}{2}}\right)$ та $\left(x^{кр}_{прав.\frac{\alpha}{2}}; +\infty\right)$, де точки $x^{кр}_{прав.\frac{\alpha}{2}}$ та $x^{кр}_{лів.\frac{\alpha}{2}}$ визначають з умов $P\left(\varphi < x^{кр}_{лів.\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$ та $P\left(\varphi > x^{кр}_{прав.\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$, їх називаються **двобічними критичними точками. Критичний простір та рівень істотності розподілені симетрично відносно осі у**.
- **Правобічна критична область** – якщо альтернативну гіпотезу сформульовано у такому вигляді:
  H_1: параметр більший від закладеного числа. Тоді критична область складається з одного інтервалу $\left(x^{кр}_{прав.\alpha}; +\infty\right)$, де точка $x^{кр}_{прав.\alpha}$ визначається з умови $P\left(\varphi > x^{кр}_{прав.\alpha}\right) = \alpha$ і називається **правобічною критичною точкою, що є відповідним рівню значущості α. За правобічного критичного простору рівень істотності α розміщується з правого боку від осі *у***;
- **Лівобічна критична область** – якщо альтернативну гіпотезу сформульовано так:
  H_1: параметр є меншим від закладеного числа. Тоді критична область складається з **одного** інтервалу: $\left(-\infty; x^{кр}_{лів.\alpha}\right)$, у якому точку $x^{кр}_{лів.\alpha}$ визначають з умови $P\left(\varphi < x^{кр}_{лів.\alpha}\right) = \alpha$ і називають **лівобічною критичною точкою, відповідною рівню значущості α. За лівобічного критичного простору рівень істотності α розміщується з лівого боку від осі *у***.



а *б* *в*

Рис. 2.1. Види розміщення критичної області:
а – лівобічна, *б* – двобічна, *в* – правобічна

Етапи верифікації гіпотез (рис. 2.2) складаються з таких дій:

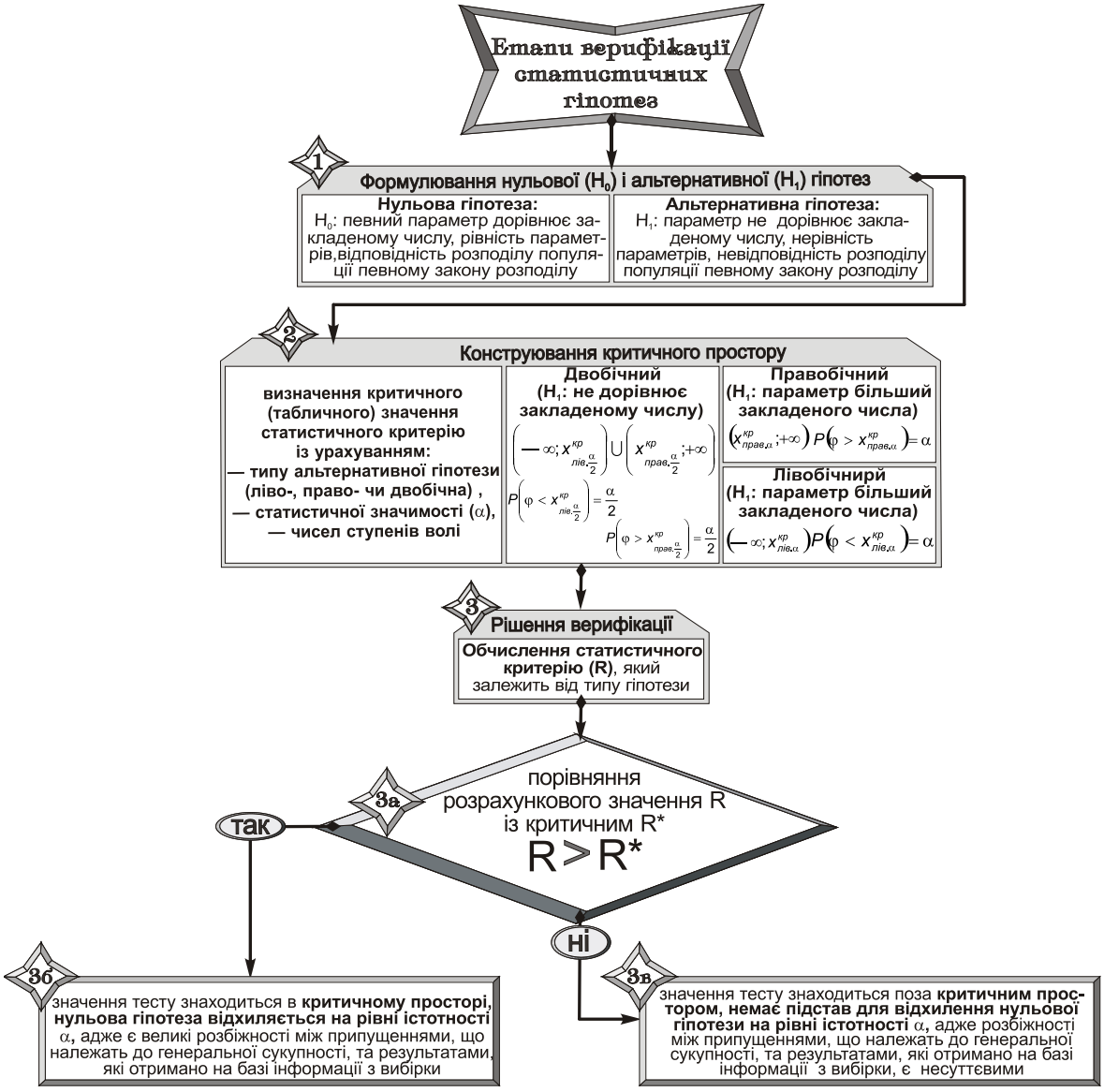
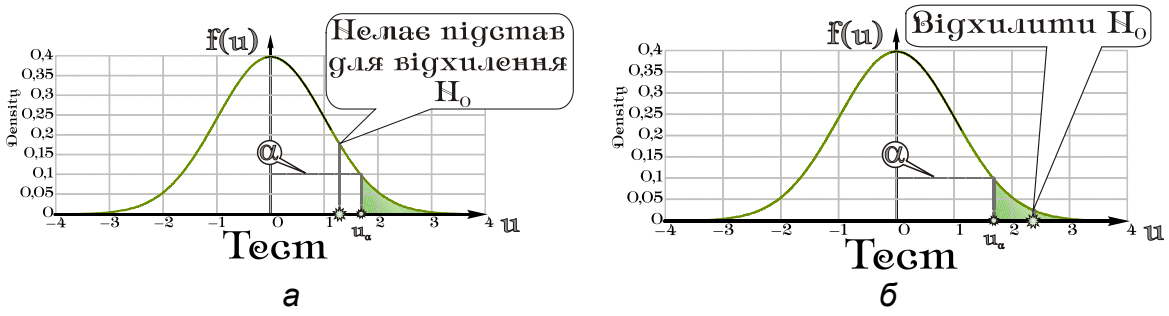


Рис. 2.2. Процедура верифікації статистичних гіпотез

- формують два види гіпотез, H_0 тестують, H_1 приймають, коли відхиллено H_0;
- визначають принципи верифікації гіпотези та виконують статистичний тест (або статистику з вибірки);
- виконують верифікацію гіпотез за допомогою тестів:
  - для параметричних гіпотез – тести істотності. Тести істотності дорівнюють відношенню різниці конкретного значення статистики з вибірки та значення перевірюваного параметра до стандартного відхилення розкладу. Тест істотності є таким правилом просування, що дає змогу ухвалити рішення про відхилення нульової гіпотези або

стверджувати, що висновки, отримані у вибірці, не дають підстав для її відхилення. У тесті істотності використовують тільки другий рядок таблиці, бо перший містить імовірнісні висновки про брак підстав для відхилення нульової гіпотези. Основне значення в тестах істотності має імовірність помилки першого виду, що повинна бути якнайменшою, тому рівень істотності (рівень значущості α) беруть 0,1; 0,05; 0,02; 0,01; 0,001;

  - для непараметричних гіпотез – тести відповідності;
- конструювання критичного простору (простору відхилень) – такого простору, у якому за наявності значення тесту відхиляють нульову гіпотезу;
- рішення верифікації (статистичний висновок), що залежить від відношення між критичним простором і значенням тесту (рис. 2.3):
  - коли значення тесту знаходиться в критичному просторі, нульову гіпотезу відхиляють на рівні істотності α, це означає, що справжньою є прийнята альтернативна гіпотеза. Така ситуація виникає, коли є великі розбіжності між припущеннями, що належать до генеральної сукупності, та результатами, отримані на основі інформації з вибірки (або між значеннями параметра, який перевіряють, та конкретними значеннями статистики з вибірки);
  - коли значення тесту знаходиться поза критичним простором, немає підстав для відхилення нульової гіпотези на рівні істотності α, це означає, що нульова гіпотеза може бути справжньою. Виявлено малі розбіжності між припущеннями, що належать до генеральної сукупності та результатами вибірки.



а *б*

Рис. 2.3. Ухвалення рішення щодо верифікації на підставі відношення між критичним простором і значенням тесту:
а – нуль-гіпотезу прийнято, *б* – нуль-гіпотезу відхилено

Твердження про те, що **немає підстав для відхилення нульової гіпотези, не тотожне із схваленням нульової гіпотези** або визнанням її за правильну. Адже, застосовуючи тести істотності, рішення про прийняття нульової гіпотези ухвалюють із ризиком припуститися помилки II виду, імовірність якої у розглядуваних тестах не беруть до уваги.

Отже, ***вся процедура верифікації гіпотез підпорядкована відхиленню нульової гіпотези,*** *оскільки, ухвалюючи рішення про відхилення H_0, ми знаємо, з якою помилкою таке рішення ухвалюємо, тому що про цю помилку ми вирішуємо самі, обираючи рівень істотності.*

2.2. ПЕРЕВІРКА НЕПАРАМЕТРИЧНИХ ГІПОТЕЗ

Для перевірки функціонального виду розподілу популяції використовують тести відповідності, що належать до класу непараметричних. *Назва походить від того, що ми порівнюємо єдність емпіричного розподілу із теоретичним.* Побудова тестів потребує дотримання двох обмежень:

- вибірка повинна бути великою;
- вибірка повинна бути простою (незалежне жеребкування – кожен елемент сукупності має однакові шанси потрапити до вибірки).

Подібно до параметричних тестів, процедуру верифікації починають з висунення **нульової та альтернативної гіпотези.**

Нульова гіпотеза (нуль-гіпотеза, H_0: $F(x)= F_0(x)$)

H_0: **популяція має розклад *G* (**продуктивність праці мулярів розподілена за нормальним законом, інтенсивність поломок фарбопульта узгоджується з законом Пуассона).

Гіпотеза, що їй суперечить – **альтернативна (конкурентна,** $H_1 : F(x) \neq F_0(x)$**)**

H_1: **популяція не має розкладу *G*** (продуктивність праці мулярів розподілена не за нормальним законом, інтенсивність поломок фарбопульта не узгоджується з законом Пуассона.

Нульова гіпотеза вказує на те, що форму розподілу в популяції ***F(x)*** *можна описати теоретичним визначенням* ***$F_0(x)$****, але альтернативна гіпотеза заперечує це твердження. Якщо досліджувана змінна має відповідний (або наближений) розподіл до закладеного, то емпіричний розподіл, визначений на основі вибірки, не повинен відрізнятися від теоретичного (гіпотетичного). Отже, за допомогою тесту відповідності досліджують розходження між гіпотетичним та емпіричним розподілами. Якщо виявлено великі розбіжності, то слід припустити, що розподіл ознаки не тотожний із закладеним (гіпотетичним) розподілом імовірності.*

Найчастіше як тест відповідності застосовують тест χ^2 – хі-квадрат (критерій Пірсона):

$$\chi^2 = \sum_{j=1}^{k} \frac{\left(f_j - f_j'\right)^2}{f_j'} = \sum_{j=1}^{k} \frac{\left(O_j - E_j\right)^2}{E_j}, \qquad (2.1)$$

де k – кількість частот (груп частот) в емпіричному розподілі (***не повинна бути меншою, ніж 5, k>=5);***

f_j (O_j) – кількість одиниць, яку має ***j***-та частота (для дискретного розподілу) або частота j-го інтервалу (для безперервного розподілу), вважають, що ***f_j>=8 (f_j>=5, інакше слід зменшити кількість частот) observed:***

$$\sum_{j=1}^{k} f_j = n,\ n \geq 50;$$

f_j‘ (E_j) – розрахункова кількість одиниць, яку має j-та теоретична частота, розраховувана (***evaluated***) згідно з формулою ***f_j=n•p_j***:

де p_i – імовірність настання визначеного числа (потрапляння варіанти до визначеного інтервалу), встановлена відповідно до гіпотетичного закону розподілу.

Статистика χ^2 **під час закладення справжньої нульової гіпотези має розподіл (розклад) χ^2 з кількістю ступенів вільності** $v=k-r-1$**,** де r – кількість оцінених параметрів на підставі вибірки (для нормального або рівномірного розподілу ***r=2***, для розподілу Пуассона, біноміального ***r=1***).

Критичний простір завжди правобічний і визначуваний через відношення:

$$P\left(\chi^2 \geq \chi^{2*}_{\alpha, v=k-r-1}\right) = \alpha.$$

Фактичне (розрахункове) значення χ^2 порівнюють із **критичним (табличним)** $\chi^{2*}_{\alpha;\nu}$: якщо табличне перевищить розрахункове $\chi^2 < \chi^{2*}_{\alpha;\nu}$, то із заданою надійністю ***P*** можна вважати, що прийнятий закон розподілу узгоджується із законом розподілу ознаки статистичної сукупності. Отже, немає сенсу відхиляти гіпотезу ***H_0***.

Якщо $\chi^2 > \chi^{2*}_{\alpha;\nu} \Rightarrow$ ***~~H_0~~*** тобто якщо значення статистики з вибірки потрапить в критичну область, то різниця між емпіричним та теоретичним розподілом, яке вимірюється різницею між теоретичними та емпіричними частотами є статистично значущою, отже, **нульову гіпотезу слід відхилити.** Тому символи H_0 перекреслено ($\text{~~}H_0\text{~~}$).

> ***Приклад*** *2.1.* Перевіримо на рівні значущості $\alpha = 0{,}01$, чи відповідний щоденний попит у будні на клей для шпалер у магазинчику «Будматеріали», що знаходиться напроти університету, **нормальному закону із математичним очкуванням 20 кг на день та стандартним відхиленням 5 кг** на підставі даних про обсяги продажу у березні – квітні (табл. 2.2.).

Нуль-гіпотезу сформулюємо так:

H_0: щоденний попит на шпалерний клей у робочі дні для торговельної точки розподілений за нормальним законом **N(x;m=20; σ=5)**.

Альтернативна гіпотеза матиме такий вигляд:

H_1: щоденний попит на шпалерний клей у робочі дні для торговельної точки не є відповідним нормальному закону **N(x;m=20; σ=5)**.

Таблиця 2.2

Вихідні дані для перевірки гіпотези про закон розподілу попиту на зошити

Дата	**15.3**	**16.3**	**17.3**	**18.3**	**19.3**	**22.3**	**23.3**	**24.3**	**25.3**	**26.3**	**29.3**	**30.3**
Обсяг продажів	15	17	15	13	26	20	19	14	30	9	14	27
Група частот	2	3	2	2	5	3	3	2	6	1	2	5

Дата	**31.3**	**1.4**	**2.4**	**5.4**	**6.4**	**7.4**	**8.4**	**9.4**	**12.4**	**13.4**	**14.4**	**15.4**	**16.4**
Обсяг продажів	20	22	23	14	20	23	17	13	19	21	22	15	17
Група частот	3	4	4	2	3	4	3	2	3	4	4	2	3

У досліджуваному періоді (вибірці) було 25 робочих днів, тому кількість спостережень **n=25.** Максимальний обсяг продажів, 30 кг клею, спостерігався 25 березня, мінімальний, 9 кг, спостерігався напередодні, тобто 25 березня.

Спочатку згрупуємо дані про обсяги продажу у шістьох категоріях (адже, відповідно до формули Стерджеса (1.24), **k=1+ln(25)/ln(2)$\approx$6).** Оскільки k=6<5, кількість груп частот є достатньою для застосування критерію Пірсона. Крок інтервалів дорівнює **4** ($\approx$3,5=(30—9)/6), подальші розрахунки зводимо у табл. 2.3.

Для обчислення теоретичних частот (E_j, наведені у графі 8 табл. 2.3) потрібні теоретичні частки. Вони, як показано у графі 7 табл. 2.3, визначають якою є різниця між суміжними теоретичними кумулятивними частотами, тобто яким є збільшення показників графи 6. Звичайно для першої групи частот теоретична частота збігається з емпіричною, тоді як показники графи 6 – теоретичні кумулятивні частоти – визначають за допомогою таблиць кумулятивного стандартного нормального розподілу або за допомогою функції MS Excel NORM.S.DIST з категорії «статистичні». Аргументом функції, який потрібен і для використання таблиць, є міра відхилення правої межі кожної групи частот від гіпотетичного математичного очікування, виражена у кількості стандартних відхилень (сигм): $Z = \frac{x - m}{\sigma}$. Розрахунок **Z** для кожної з груп частот наведено в графі 5 табл. 2.3. В останній графі табл.2.3 (графа 10) надано проміжні розрахунки безпосередньо критерію

Пірсона, тобто сума показників графи 10 і є **розрахунковим значенням критерію Хі-квадрат:**

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} = \frac{(1-1)^2}{1} + \frac{(8-4)^2}{4} + \frac{(8-7)^2}{7} + \frac{(5-7)^2}{7} + \frac{(2-4)^2}{4} + \frac{(1-1)^2}{1} =$$

$$= 0 + 4 + 0{,}14 + 1{,}57 + 1{,}0 = 5{,}71.$$

Таблиця 2.3

Допоміжна розрахункова таблиця для визначення критерію Пірсона (χ^2 — хі-квадрат)

Група частот	Межі попиту, шт.	Права межа	Кількість спостережень, емпірична частота, O_j	Кількість сигм для правої межі $Z = \frac{x-m}{\sigma}$	Теоретична кумулятивна частка (функція кумулятивного нормального розподілу Ф(Z)	Теоретична частка = Δ кумулятивна частка	Теоретична (розрахункова) частота, E_j	Різниця між емпіричною і теоретичною частотами, O_j-E_j	$\frac{(O_j-E_j)^2}{E_j}$
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6 =NORM.S.DIST([5];1)	**7**= Δ [6]	**8**=[7]•n	**9**=[4]-[8]	**10**= $[9]^2/[8]$
1	до 12	12	1	-1,6	**0,055**=NORM.S.DIST (-1,6;1)	**0,055**=0,055	1=0,055• 25	**0**=1-1	**0**=$0^2/1$
2	понад 12 - 16	16,0	8	-0,8	**0,212**=NORM.S.DIST (-0,8;1)	**0,157**=0,212-0,055	4=0,157• 25	**4**=8-4	**4**=$4^2/4$
3	понад 16 - 20	20,0	8	0	**0,5**=NORM.S.DIST(0;1)	**0,288**=0,5-0,212	7=0,288• 25	**1**=8-7	**0,14**= $1^2/7$
4	понад 20 - 24	24,0	5	0,8	**0,788**=NORM.S.DIST (0,8;1), або =1-NORM.S.DIST (-0,8;1)	**0,288**=0,788-0,5	7=0,288• 25	**-2**=5-7	**0,57**= $(-2)^2/7$
5	понад 24 - 28	28,0	2	1,6	**0,945**=NORM.S.DIST (1,6;1), або 1-=NORM.S.DIST (-1,6;1)	**0,157**= 0,945-0,788	4= 0,157•25	**-2**=2-4	**1**= $(-2)^2/4$
6	понад 28	∞	1	3,5	**1**=NORM.S.DIST(3,5;1)	**0,055**= 1-0,945	1= 0,055•25	**0**=1—1	**0**= $0^2/1$

Для верифікації нуль-гіпотези про нормальний закон розподілу попиту **N(x;m=20;σ=5)** потрібне **т**абличне значення $\chi^{2*}_{0,01;3}$ **з кількістю ступенів вільності** *v=6–2–1=3*, адже обґрунтований рівень значущості $\alpha = 0{,}01$, а спостереження вибірки розподілено на **шість категорій** і при цьому кількість параметрів нормального розподілу **r=2**: математичне очікування та стандартне відхилення. Його можна визначити за допомогою статистичних таблиць, наприклад, табл. 2.7, або із застосуванням функції MS Excel ХИ2ОБР, а в сучасних англомовних версіях табличних процесорів CHISQ.INV.RT з категорії **статистичних**, а саме: =ХИ2ОБР(0.01;6-2-1), чи

=CHISQ.INV.RT(0.01;6-2-1). Таким чином, $\chi^{2*}_{0,01;3} = 11{,}345$. Оскільки **розрахункове значення χ^2 виявилось меншим за табличне** $\left(5{,}71 < 11{,}345 \Rightarrow \chi^2 < \chi^{2*}_{0,01;3} \Rightarrow H_0(i)\right)$, то із заданою надійністю **P=0,99** можна вважати, що прийнятий закон розподілу узгоджується із нормальним законом розподілу ознаки статистичної сукупності. Отже, немає сенсу відхиляти нуль-гіпотезу H_0 щодо нормального закону попиту шпалерного клею у робочі дні із математичним очікуванням 20 кг на день та стандартним відхиленням 5 кг на день. При цьому помилка у визначенні закону розподілу попиту може бути лише в одному випадку зі 100.

2.3. ПЕРЕВІРКА ПАРАМЕТРИЧНИХ ГІПОТЕЗ

2.3.1. Основні теоретичні відомості

Параметричні гіпотези стосуються параметрів, **що характеризують популяцію**. Насамперед перевіряють очікувані значення таких параметрів:

- значення середніх величин в одній чи кількох (найчастіше двох) сукупностях;
- відносних часток (також в одній чи двох сукупностях), що часто називають імовірностями;
- значення дисперсій однієї, двох або кількох генеральних сукупностей.

З цією метою застосовують низку статистичних критеріїв: *Z*-критерій нормального розподілу, *t*-критерій розподілу Стьюдента, *F*-критерій Фішера, Хі-квадрат (χ^2) розподілу Пірсона. *Z*-критерій нормального розподілу, *t*-критерій розподілу Стьюдента застосовують для перевірки гіпотез стосовно середніх значень і часток, причому *t*-критерій потрібен для дослідження малих вибірок, обсяг яких менший за 30 спостережень (*n*<30). Для верифікації гіпотез відносно дисперсій застосовують статистичні критерії Фішера та Пірсона. Основні випадки застосування верифікації гіпотез та необхідні для цього статистичні критерії зведено в табл. 2.4.

Таблиця 2.4

Основні випадки верифікації параметричних гіпотез

Параметр сукупності, очікуване значення якого перевіряють	**Кількість сукупностей**	**Розмір вибірок**	**Статистичний критерій**	**Критичне значення**	**Шифр типу гіпотез для подальшого викладу матеріалу**
1	2	3	4	5	6
Середня величина	одна	велика	$Z = \frac{\widetilde{x} - \overline{x}_{H_0}}{S} \bullet \sqrt{n}$	Z^*_α	Тип «А»

Параметр сукупності, очікуване значення якого перевіряють	**Кількість сукупностей**	**Розмір вибірок**	**Статистичний критерій**	**Критичне значення**	**Шифр типу гіпотез для подальшого викладу матеріалу**
		мала	$t = \frac{\tilde{x} - \bar{x}_{H_0}}{S} \cdot \sqrt{n-1}$	$t^*_{\alpha;\nu=n-1}$	
	дві	великі	$Z = \frac{\tilde{x}_1 - \tilde{x}_2 - \left(\bar{x}_{1H_0} - \bar{x}_{2H_0}\right)}{\sqrt{\frac{S_1^2(x)}{n_1} + \frac{S_2^2(x)}{n_2}}}$	Z^*_α	
		малі	$t = \frac{\tilde{x}_1 - \tilde{x}_2 - \left(\bar{x}_{1H_0} - \bar{x}_{2H_0}\right)}{\sqrt{\frac{n_1 \cdot S_1^2(x) + n_2 \cdot S_2^2(x)}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$t^*_{\alpha;\nu=n_1+n_2-2}$	

Закінчення табл. 2.4

1	2	3	4	5	6
Середня величина	три та більше, (***m***)	не має значення, бажано однакові n_j=const	$F = \frac{\frac{\sum_{j=1}^{m}\left(\bar{x}_i - \bar{x}\right)^2 \cdot n_j}{m-1}}{\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2}{n-m}}$	$F^*_{\alpha;\nu_1=m-1;\nu_2=n-m}$	Тип «В»
Відносна частка (імовірність)	одна	велика n_j=const	$Z = \frac{w - \bar{p}_{H_0}}{\sqrt{\frac{w \cdot (1-w)}{n}}}$	Z^*_α	Тип «Б»
	дві	великі, можна різні, $n_1 \neq n_2$	$Z = \frac{w_1 - w_2}{\sqrt{\bar{p} \cdot (1-\bar{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, $\bar{p} = \frac{m_1 + m_2}{n_1 + n_2}$		
Дисперсія	одна	мала	$\chi^2 = \frac{n \cdot S^2(x)}{\sigma^2_{H_0}}$	$\left(\chi^2_{\alpha;\nu=n-1}\right)^*$	Тип «В»
	дві	малі	$S = \frac{\frac{n_1 \cdot S_1^2(x)}{n_1 - 1}}{\frac{n_2 \cdot S_2^2(x)}{n_2 - 1}}$	$F_{\alpha;\nu_1=n_1-1;\nu_2=n_2-1}{}^*$	

Розглянемо докладніше процедуру верифікації гіпотез кожного з трьох типів.

2.3.2. ГІПОТЕЗИ ЩОДО СЕРЕДНЬОГО ЗНАЧЕННЯ ВЕЛИЧИН (ТИП «А»)

Випадок одної сукупності. Велика вибірка, *n*>30

У такому випадку **нульову гіпотезу** формулюють так: ***H₀: «***$\bar{X} = \bar{X}_{H_0}$», тобто ***H₀: «***Середнє значення в деякій сукупності дорівнює певному математичному сподіванню величини».

Альтернативна гіпотеза залежно від сутності проблеми може бути**:**

- **двобічною: *H₁: «***$\bar{X} \neq \bar{X}_{H_0}$», тобто ***H₁: «***Середнє значення в деякій сукупності **не дорівнює** певному математичному сподіванню величини»;
- **лівобічною: *H₁: «***$\bar{X} < \bar{X}_{H_0}$», тобто ***H₁: «***Середнє значення в деякій сукупності є **меншим** за певне математичного сподівання величини»;
- **правобічною: *H₁: «***$\bar{X} > \bar{X}_{H_0}$», тобто ***H₁: «***Середнє значення в деякій сукупності є **більшим** за певне математичне сподівання величини».

Верифікацію гіпотез виконують за допомогою *Z*-тесту:

$$Z = \frac{\tilde{X} - \bar{X}_{H_0}}{S_{\bar{x}}} = \frac{\tilde{X} - \bar{X}_{H_0}}{\frac{S}{\sqrt{n}}} = \frac{\tilde{X} - \bar{X}_{H_0}}{S} \bullet \sqrt{n}\,, \qquad (2.2)$$

де $\tilde{X}$ та $\bar{X}_{H_0}$ – середні значення відповідно за вибіркою та за гіпотетичним (H_0) припущенням;

$S_{\bar{x}}$, S – відповідно стандартне відхилення середнього та стандартне відхилення розподілу з вибірки;

N – обсяг вибірки.

Якщо значення тесту **перевищить табличне (критичне)** значення **статистики нормального розподілу** Z^*_α за обраного дослідником рівня істотності α**, нульову гіпотезу про рівність середнього значення** генеральної сукупності очікуваному значенню середньої величини **слід спростувати та взяти альтернативну:**

$$|Z| > |Z^*_\alpha| \Rightarrow \text{H}_0 \Rightarrow \text{H}_1(\text{i}),$$

При цьому залишається ймовірність ухвалення помилкового рішення на рівні істотності α. Інакше, коли значення Z-статистики виявиться меншим за табличне за рівня істотності α**, нульову гіпотезу про рівність середнього значення** генеральної сукупності очікуваному значення варто прийняти: $|Z| \le Z^*_\alpha \Rightarrow \text{H}_0\ (\text{i})$.

Критичне значення Z^*_α визначають за допомогою статистичних таблиць кумулятивної функції нормального розподілу, зважаючи на тип критичного простору:

– **для двобічного** критичного простору – як розв'язок рівняння $P\left(|Z| \geq Z_{\alpha}^{*}\right) = \frac{\alpha}{2}$. При цьому **нуль-гіпотеза** має містити знак **суворої рівності («=»)**, тоді як **альтернативна** – знак «$\neq$»;

– **для однобічного** критичного простору – як розв'язок рівняння $P\left(|Z| \geq Z_{\alpha}^{*}\right) = \alpha$. Звичайно, тоді **нуль-гіпотеза** має містити знак **несуворої нерівності** («$\leq$»,або «$\geq$»), **альтернативна** – знак **суворої нерівності** (відповідно «>», або «<»).

Найчастіше для верифікації гіпотез щодо математичного очікування ознаки соціально-економічних явищ беруть рівні істотності $\alpha = 0{,}1$, чи 0,05, чи 0,025, чи 0,02, чи 0,01. Критичні значення Z_{α}^{*} в разі використання нормального розподілу за таких імовірностей помилкових рішень наведено у табл. 2.5.

У медико-біологічних дослідженнях помилкове рішення має особливо небезпечні наслідки, рівень істотності є значно нижчим – часто він не перевищує однієї тисячної $\alpha = 0{,}001$.

Таблиця 2.5

Критичні значення Z_{α}^{*}

Характеристика критичного простору	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}02$	$\alpha = 0{,}01$
двобічний	1,64	1,96	2,24	2,33	2,58
правобічний	1,28	1,64	1,96	2,05	2,33
лівобічний	-1,28	-1,64	-1,96	-2,05	-2,33

Приклад 2.2. Розглянемо приклад верифікації гіпотез про рівність середнього для великої вибірки. На рівні істотності $\alpha = 0{,}01$ перевіряють гіпотезу про те, що денна вартість оренди деякого будівельного механізму становить **8 500 грн (**$\overline{x}_{H_0} = 8500$**)**. З цією метою опитано **169 орендодавців (n=169)**. За результатами опитування, середня вартість одного дня оренди – **8 610 грн** ($\widetilde{x} = 8610$) зі стандартним відхиленням у вибірці **520 грн/день** (S=520 грн/день, відповідно вибіркова дисперсій становить S^2=270400 (грн/день)^2).

Етап 1. Формулювання гіпотез

Нуль-гіпотеза: H_0: «В середньому денна вартість оренди техніки становить 8 500 грн, тобто $\overline{x} = 8500$ ».

Альтернативна гіпотеза: H_1: «Середня вартість оренди такої техніки становить не 8 500 грн/день, тобто $\overline{x} \neq 8500$ ».

Критичний простір гіпотези **двобічний:** в записі альтернативної гіпотези є знак «$\neq$», тобто насправді у генеральній сукупності вартість денної оренди може бути як вищою, так і нижчою за очікуване, гіпотетичне значення.

Етап 2. Розрахунок Z-статистики

$$Z = \frac{\tilde{x} - \bar{x}_{H_0}}{S} \cdot \sqrt{n} = \frac{8610 - 8500}{520} \cdot \sqrt{169} = 2{,}75 .$$

Оскільки обсяг вибірки перевищує 30 одиниць (n=169>30), вибірка є великою, а тому можна застосовувати Z-статистику (2.2).

Етап 3. Конструювання критичного простору для двобічної критичної області

Зі статистичної таблиці критичних значень Z^*_α (табл. 2.5) $Z^*_{\frac{\alpha}{2}=\frac{0{,}01}{2}} = Z^*_{0{,}005} = 2{,}58$. Той самий результат можна отримати за допомогою функції НОРМСТОБР чи NORM.S.INV з категорії «статистичні» MS Excel, тобто вводячи до вільної клітинки формулу «=НОРМСТРОБР(0,01/2)» чи «=NORM.S.INV(1-0,01/2)»:

$$\left(-\infty; x^*_{лів\,\frac{\alpha}{2}}\right) \cup \left(x^*_{прав\,\frac{\alpha}{2}}; +\infty\right) \Rightarrow (-\infty; -2{,}58) \cup (2{,}58; +\infty).$$

Етап 4. Розв'язок верифікації: $|-2{,}75| > |2{,}58| \Rightarrow \cancel{H_0} \Rightarrow H_1(i)$. Це дає підставу **спростувати** нульову гіпотезу про рівність денної ставки оренди досліджуваного виду будівельної техніки очікуваному значенню 8500 грн, вважаючи імовірність помилкового рішення на рівні $\alpha = 0{,}01$. Графічно значення тесту потрапляє до критичної області (рис. 2.4), що свідчить про досить велику розбіжність між даними щодо очікуваної орендної ставки досліджуваних будівельних механізмів та результатом, отриманим у вибірці.

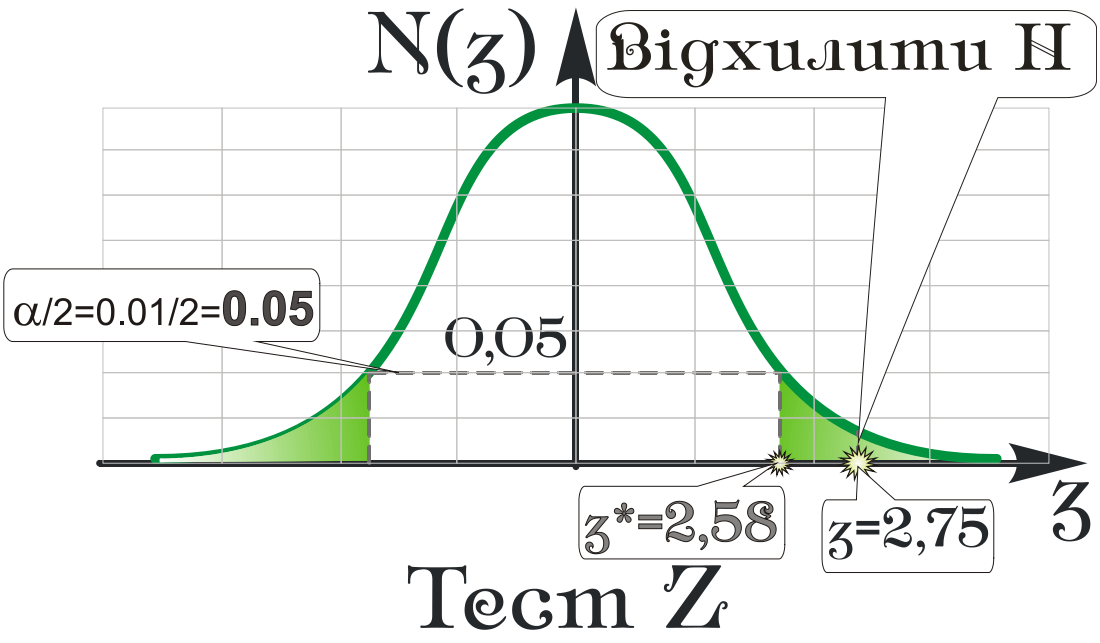


Рис. 2.4. Відхилення нульової гіпотези на підставі відношення
між критичним простором і значенням Z-тесту

Якби на основі тих самих вибіркових даних потрібно було б перевірити гіпотезу про те, що вартість денної оренди певного виду будівельної техніки **не перевищує 8 500 грн, етапи верифікації** були б такими:

Етап 1. Формулювання гіпотез

Нуль-гіпотеза: H_0: «Вартість денної оренди будівельного механізму становить 8500 грн, тобто $\bar{x} = 8500$ ».

Альтернативна гіпотеза: H_1: «Вартість денної оренди будівельного механізму перевищує 8500 грн, тобто $\bar{x} > 8500$ ».

Критичний простір гіпотези тепер **правобічний** (отже, **однобічний:** в запису альтернативної гіпотези є знак суворої нерівності «>», тобто насправді, у генеральній сукупності денна вартість оренди може бути лише вищою за очікуване, гіпотетичне значення.

Етап 2. Розрахунок Z-статистики збігається з попереднім випадком, тобто

$$Z = \frac{\tilde{x} - \bar{x}_{H_0}}{S} \cdot \sqrt{n} = \frac{8610 - 8500}{520} \cdot \sqrt{169} = 2{,}75 .$$

Оскільки обсяг вибірки перевищує 30 одиниць (n=169>30), вибірка є великою, а тому може бути застосована Z-статистика (2.2).

Етап 3. Конструювання критичного простору для двобічної критичної області

Зі статистичної таблиці критичних значень Z^*_α (див. табл. 2.5.) $Z^*_{\alpha=0{,}01} = Z^*_{0{,}01} = 2{,}33$. Такий самий результат можна отримати за допомогою функції НОРМСТОБР з категорії «статистичні» MS Excel із суперпозицією за функцією ABS (для модуля значення), тобто вводячи до вільної клітинки формулу «=ABS(НОРМСТРОБР(0,01))». В пізніх англомовних версіях процесора слід ввести «**=NORM.S.INV(1-0.01)**». Оскільки критична область є однобічною, імовірність помилки «зосереджена» з одного краю кривої розподілу, тому аргумент функції **NORM.S.INV** не містить ділення величини довірчої імовірності на 2. Одержимо область

$$\left(x^*_\alpha; +\infty\right) \Rightarrow (2{,}33; +\infty).$$

Етап 4. Рішення верифікації: $|-2{,}75| > |2{,}33| \Rightarrow \not{H}_0 \Rightarrow H_1(i)$. Це дає нам підставу **спростувати** нульову гіпотезу й у цьому випадку. Тобто на рівні істотності $\alpha = 0{,}01$ не можна стверджувати, що вартість оренди досліджуваного виду будівельної техніки не перевищить очікуваної межі у 8 500 грн/день. Графічна інтерпретація значення тесту буде аналогічною зображеному на рис. 2.4.

Випадок однієї сукупності. Мала вибірка, коли n<30

У такому випадку верифікацію гіпотез виконують за допомогою *t*-тесту (*t*-статистики):

$$t = \frac{\widetilde{x} - \bar{x}_{H_0}}{S_{\bar{x}}} = \frac{\widetilde{x} - \bar{x}_{H_0}}{\frac{S}{\sqrt{n-1}}} = \frac{\widetilde{x} - \bar{x}_{H_0}}{S} \cdot \sqrt{n-1}, \qquad (2.3)$$

де $\widetilde{x}$ та $\bar{x}_{H_0}$ – середні значення відповідно у вибірці та за гіпотетичним припущенням;

$S_{\bar{x}}$, S – відповідно стандартне відхилення середнього та стандартне відхилення розподілу з вибірки;

n – обсяг вибірки.

Якщо значення тесту **перевищить табличне (критичне)** значення **статистики розподілу Стьюдента** $t^*_{\alpha;\nu}$ за обраного дослідником рівня істотності α та кількості ступенів вільності, рівної $\nu = n-1$ **, нульову гіпотезу про рівність середнього значення** генеральній сукупності очікуваному значенню середньої величини **слід спростувати та прийняти альтернативну:**

$$|t| > |t^*_{\alpha;\nu}| \Rightarrow \text{H}_0 \Rightarrow \text{H}_1(\text{i}),$$

При цьому залишається імовірність ухвалення помилкового рішення на рівні істотності α**.** В іншому випадку, коли значення t-статистики виявиться меншим за табличне за рівня істотності α та кількості ступенів вільності *n* – 1**, нульову гіпотезу про рівність середнього значення** генеральній сукупності очікуваному значенню варто взяти: $|t| \le t^*_{\alpha;\nu} \Rightarrow \text{H}_0$ (i) .

Критичне значення $t^*_{\alpha;\nu}$ визначають за допомогою статистичних таблиць функції розподілу Стьюдента відповідно не лише до типу критичного простору (заголовки стовпців), але й виходячи з кількості ступенів вільності (заголовки рядків). Крім того, також для верифікації гіпотез щодо математичного очікування ознаки соціально-економічних явищ беруть рівні істотності $\alpha = 0{,}1$, чи 0,05, чи 0,025, чи 0,02,чи 0,01. Фрагмент таблиці функції розподілу Стьюдента подано у табл. 2.6.

Таблиця 2.6

Критичні значення $t^*_{\alpha;\nu}$

ν	**Двобічна критична область**						ν	**Двобічна критична область**					
	$\alpha = 0{,}2$	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}02$	$\alpha = 0{,}01$		$\alpha = 0{,}2$	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}02$	$\alpha = 0{,}01$
	Однобічна критична область							**Однобічна критична область**					
	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}0125$	$\alpha = 0{,}01$	$\alpha = 0{,}005$		$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}0125$	$\alpha = 0{,}01$	$\alpha = 0{,}005$
1	3,078	6,314	12,706	25,452	31,821	63,657	16	1,337	1,746	2,120	2,473	2,583	2,921
2	1,886	2,920	4,303	6,205	6,965	9,925	17	1,333	1,740	2,110	2,458	2,567	2,898
3	1,638	2,353	3,182	4,177	4,541	5,841	18	1,330	1,734	2,101	2,445	2,552	2,878
4	1,533	2,132	2,776	3,495	3,747	4,604	19	1,328	1,729	2,093	2,433	2,539	2,861
5	1,476	2,015	2,571	3,163	3,365	4,032	20	1,325	1,725	2,086	2,423	2,528	2,845
6	1,440	1,943	2,447	2,969	3,143	3,707	21	1,323	1,721	2,080	2,414	2,518	2,831

7	1,415	1,895	2,365	2,841	2,998	3,499	22	1,321	1,717	2,074	2,405	2,508	2,819
8	1,397	1,860	2,306	2,752	2,896	3,355	23	1,319	1,714	2,069	2,398	2,500	2,807
9	1,383	1,833	2,262	2,685	2,821	3,250	24	1,318	1,711	2,064	2,391	2,492	2,797
10	1,372	1,812	2,228	2,634	2,764	3,169	25	1,316	1,708	2,060	2,385	2,485	2,787
11	1,363	1,796	2,201	2,593	2,718	3,106	26	1,315	1,706	2,056	2,379	2,479	2,779
12	1,356	1,782	2,179	2,560	2,681	3,055	27	1,314	1,703	2,052	2,373	2,473	2,771
13	1,350	1,771	2,160	2,533	2,650	3,012	28	1,313	1,701	2,048	2,368	2,467	2,763
14	1,345	1,761	2,145	2,510	2,624	2,977	29	1,311	1,699	2,045	2,364	2,462	2,756
15	1,341	1,753	2,131	2,490	2,602	2,947	30	1,310	1,697	2,042	2,360	2,457	2,750

Приклад 2.3. Розглянемо приклад верифікації гіпотез щодо рівності середнього для малої вибірки. *Виробник тонера деякої марки обіцяє гарантійний строк служби картриджа – **не менший за 5 000 копій** ($\bar{x}_{H_0} = 5000$**).** З вибірки, що складається з **17 картриджів (n=17**<30**)** було отримане середнє значення строку служби – **5 005 копій** ($\bar{x} = 5005$) з дисперсією у вибірці **100 копій^2** (S^2=100 (копій)^2, відповідно стандартне відхилення у вибірці становить S=10 копій). Потрібно встановити, чи є правдивою обіцянка виробника на рівні значущості $\alpha = 0{,}01$.*

Етап 1. Формулювання гіпотез

Нуль-гіпотеза: H_0: «Очікувана потужність картриджа є не меншою, ніж 5 000 копій, тобто $\bar{x} \geq 5000$ ».

Альтернативна-гіпотеза: H_1: «Очікувана потужність картриджа є меншою, ніж 5 000 копій, тобто $\bar{x} < 5000$ ».

*Критичний простір гіпотези – **однобічний (лівобічний):** в записі альтернативної гіпотези є знак* «<», тобто *насправді у генеральній сукупності потужність картриджа може бути лише нижчою за очікуване, гіпотетичне значення.*

Етап 2. Розрахунок t-статистики:

$$t = \frac{\bar{x} - \bar{x}_{H_0}}{S} \cdot \sqrt{n - 1} = \frac{5005 - 5000}{10} \cdot \sqrt{17 - 1} = 2.$$

Оскільки обсяг вибірки не перевищує 30 одиниць (n=17<30), вибірка є великою, а тому може застосовуватись лише t-статистика (2.3).

Етап 3. Конструювання критичного простору для двобічної критичної області

Для значущості $\alpha = 0{,}01$ зі статистичної таблиці критичних значень $t^*_{\alpha;\nu}$ (табл.2.6.) $t^*_{\alpha=0{,}01;\nu=17-1} = t^*_{0{,}01;16} = 2{,}583$. Той самий результат можна отримати за допомогою функції СТЬЮДРАСПОБР з категорії «статистичні» MS Excel, тобто вводячи до вільної клітинки формулу «=СТЬЮДРАСПОБР(0,01*2;16)», адже функція видає значення *t*-критерію для **двобічної критичної області**. Одержимо область $\left(x^*_{\alpha;\nu};+\infty\right) \Rightarrow \left(2{,}583 + \infty\right)$.

У пізніших англомовних версіях Excel розроблено окрему функцію для **однобічної критичної області розподілу Стьюдента T.INV**. Особливістю цієї функції є те, що аргумент на позначення імовірності має бути заданий не як імовірність помилки α, а навпаки, **як імовірність правильної гіпотези**, тобто як $1-\alpha$. У цьому прикладі, таким чином, слід ввести в клітинку електронної таблиці вираз: **=T.INV(1-0.01;17-1)**. Він дасть результат, рівний $t^*_{(0,01;16)}$, тобто **2,83**. Та якщо задати не $1-\alpha$, а α, **програма видає від'ємне значення *t**. Отже, нехтуючи знаком результату, можна як аргумент функції **T.INV** зазначати рівень значущості гіпотези, подібно до попередніх версій.

Етап 4. Рішення верифікації

Для рівня істотності $\alpha = 0{,}01$ ***нуль-гіпотеза є істинною,*** *оскільки розрахункове значення t-критерію не потрапляє до критичного простору:* $2 < 2{,}583 \Rightarrow H_0(i)$. *Останнє дає нам підставу* ***прийняти*** *нульову гіпотезу про потужність картриджів в обсязі, не нижчому за очікуване значення – 5 000 копій, приймаючи імовірність помилкового рішення на рівні* $\alpha = 0{,}01$. *Оскільки значення t-тесту не потрапляє до критичної області (рис. 2.5), тобто не слід відхиляти нульову гіпотезу, спростовуючи обіцянку виробника картриджів. При цьому помилка може бути лише в одному випадку зі 100.*

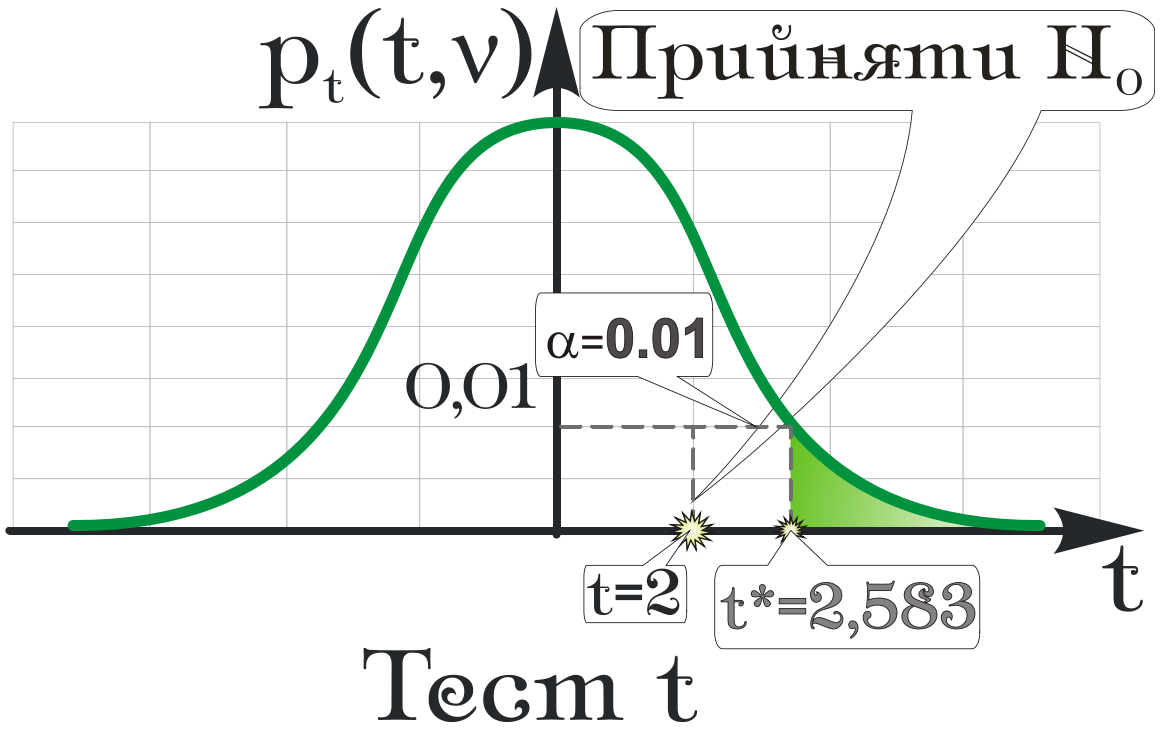


Рис. 2.5. Ухвалення рішення про істинність нульової гіпотези
на основі відношення між критичним простором і значенням t-тесту

Випадок двох сукупностей. Велика (*n*>30) та мала (*n*<30) вибірки

Дуже часто у практичних дослідженнях доводиться порівнювати дві або більше популяцій (сукупностей) з погляду рівності певного параметру. Найчастіше порівнюють дві сукупності з приводу однаковості їхніх середніх, частостей (імовірностей) прояву ознак, чи дисперсій, тобто мінливостей ознак.

Перевіряючи гіпотезу про рівність двох значень певної характеристики очікуваним величинам відповідно $\bar{x}_{1H_0}$ та $\bar{x}_{2H_0}$, нульову гіпотезу можна сформулювати за такими двома варіантами:

$$H_0: \bar{x}_{1H_0} = \bar{x}_{2H_0} \text{ або } H_0: \bar{x}_{1H_0} - \bar{x}_{2H_0} = 0.$$

Відповідно до сутності досліджуваної проблеми, кожен з трьох можливих варіантів альтернативної гіпотези (ліво-, право- та двобічна) також може бути виражений двома способами:

- двобічна: $H_1: \bar{x}_1 \neq \bar{x}_2$ або $H_1: \left(\bar{x}_1 - \bar{x}_2\right)_{H_0} \neq 0$;
- правобічна: $H_1: \bar{x}_1 > \bar{x}_2$ або $H_1: \left(\bar{x}_1 - \bar{x}_2\right)_{H_0} > 0$;
- лівобічна: $H_1: \bar{x}_1 < \bar{x}_2$ або $H_1: \left(\bar{x}_1 - \bar{x}_2\right)_{H_0} < 0$.

Залежно від обсягу вибірки, за даними якої виконують верифікацію, можна застосовувати:

- Z-статистику нормального розподілу (2.4) для великих вибірок, сукупний обсяг яких перевищує 120 спостережень ($n_1+n_2>120$):

$$Z = \frac{\tilde{x}_1 - \tilde{x}_2 - \left(\bar{x}_{1H_0} - \bar{x}_{2H_0}\right)}{\sqrt{\frac{S_1^2(x)}{n_1} + \frac{S_2^2(x)}{n_2}}} = \frac{\tilde{x}_1 - \tilde{x}_2}{\sqrt{\frac{S_1^2(x)}{n_1} + \frac{S_2^2(x)}{n_2}}}, \qquad (2.4)$$

адже згідно з нуль-гіпотезою $\bar{x}_{1H_0} - \bar{x}_{2H_0} = 0$;

- t-статистику Стьюдента із (2.5) кількістю ступенів вольності $\nu = n_1 + n_2 - 2$ у разі малих вибірок:

$$t = \frac{\tilde{x}_1 - \tilde{x}_2 - \left(\bar{x}_{1H_0} - \bar{x}_{2H_0}\right)}{\sqrt{\frac{n_1 \cdot S_1^2(x) + n_2 \cdot S_2^2(x)}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\tilde{x}_1 - \tilde{x}_2}{\sqrt{\frac{n_1 \cdot S_1^2(x) + n_2 \cdot S_2^2(x)}{n_1 + n_2 - 2}}} \cdot \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \qquad (2.5)$$

адже згідно з нуль-гіпотезою $\bar{x}_{1H_0} - \bar{x}_{2H_0} = 0$.

Якщо крім нуль-гіпотези, що означає рівність середніх у кожній з вибірок, обидві вони однакові за розміром, тобто ***n1=n2,*** формули (2.4) та (2.5) додатково спрощують:

$$Z = = \frac{\tilde{x}_1 - \tilde{x}_2}{\sqrt{\frac{S_1^2(x) + S_2^2(x)}{n_1}}};$$

та

$$t = \frac{\widetilde{x}_1 - \widetilde{x}_2}{\sqrt{\frac{S_1^2(x) + S_2^2(x)}{1 - \frac{1}{n_1}}}} \bullet \sqrt{n_1}\,.$$

Подальші етапи верифікації гіпотези збігаються із випадком однієї вибірки відповідно до загальної схеми перевірки гіпотез (рис. 2.2), а саме**:**

- відповідно до обраного дослідником рівня істотності визначають **критичне значення критерію (**Z або t) та конструюють критичну область;
- порівнюються розрахункове й табличне значення критеріїв;
- у разі перевищення розрахунковим критерієм табличного, а отже, і потрапляння значення статистичного тесту до критичної області, **нульову гіпотезу відкидають та беруть альтернативну**. В іншому випадку, якщо розрахункова статистика не перевищить критичного значення, нуль-гіпотезу приймають на визначеному дослідником рівні істотності.

Приклад 2.4. Розглянемо приклади верифікації гіпотез про рівність середніх у двох сукупностях для різних за розміром вибірок.

Побутує думка, ніби кріпильні металеві деталі вітчизняного виробництва дешевші за імпортні. З метою перевірки такого твердження проаналізовано прайси різних торговельних закладів м. Києва. Виявилось, що, за інформацією, наданою **50 ритейлорами,** які продають імпортні вироби (n_1=50), середня ціна за 10 шт. становить **7 000 грн** ($\widetilde{x}_1 = 7000$), причому стандартне відхилення у вибірці становить **1 700 грн** (S_1=1700 грн). Результати вивчення прайсів **100 реалізаторів** вітчизняної продукції (n_2=100) свідчать про те, що середня ціна кріпильних деталей **6 800 грн** за 10 шт. ($\widetilde{x}_2 = 6800$) із стандартним відхиленням у вибірці в розмірі **2000 грн** (S_2=2000 грн). Рівень істотності встановимо в розмірі $\alpha = 0{,}02$, а нуль-гіпотезу сформулюємо так, нібито **немає жодних розбіжностей в ціні металевих кріпильних деталей,** незалежно від країни виробництва. **Взагалі у більшості статистичних та економетричних досліджень зазвичай формулюють нуль-гіпотези, виходячи з того, що розбіжностей та зв'язків** між досліджуваними явищами, ознаками немає. Отже,

$$H_0: \overline{x}_{1H_0} = \overline{x}_{2H_0}, \text{ або } H_0: \overline{x}_{1H_0} - \overline{x}_{2H_0} = 0\,.$$

Тоді альтернативна гіпотеза є **двобічною**: **H_1:** $\overline{x}_{1H_0} \neq \overline{x}_{2H_0}$ або **H_1:** $\overline{x}_{1H_0} - \overline{x}_{2H_0} \neq 0$ (оскільки містить знак «≠»).

Зважаючи на те, що вибірки є великими: $n_1+n_2=50+100>120$, слід обчислити Z-статистику (2.4):

$$Z = \frac{\tilde{x}_1 - \tilde{x}_2}{\sqrt{\frac{S_1^2(x)}{n_1} + \frac{S_2^2(x)}{n_2}}} = \frac{7000 - 6800}{\sqrt{\frac{1700^2}{50} + \frac{2000^2}{100}}} = 0{,}64 .$$

Процентна точка для двостороннього критичного простору, розрахована із співвідношення $P\left(|Z| \geq Z_\alpha^*\right) = \frac{\alpha}{2}$, для $\alpha = 0{,}02$ становить (див. табл. 2.5.) Z*=2,33. Те саме число можна отримати за допомогою попередніх Excel, вводячи до клітинки суперпозицію функцій =ABS(НОРМСТОБР(0.02/2)). У сучасних англомовних версіях табличного процесора правильний результат дає функція **=NORM.S.INV(1-0.02/2)**. **Підрахована статистика не потрапляє до критичного простору** (як на рис. 2.3, *а*), а отже, **немає підстав для спростування нуль-гіпотези:** Z=0,64<Z*=2,33. Таким чином, думка про вартість імпортних деталей є хибною, тоді як рівень їхньої ціни не залежить від країни походження. Стверджуючи це, ми можемо помилятись лише в 2% випадків.

Наведемо приклад верифікації **однобічної** гіпотези щодо **малої** вибірки, якою стверджується розбіжність двох середніх значень певної характеристики, що дорівнює певній очікуваній величині. У такому разі нуль-гіпотезу слід сформулювати як H_0: $\bar{x}_{1H_0} - \bar{x}_{2H_0} = \Delta\bar{x}$ і відповідно H_0: $\bar{x}_{1H_0} \neq \bar{x}_{2H_0}$, або ж $\bar{x}_{1H_0} - \bar{x}_{2H_0} \neq 0$. Тоді формула (2.5) буде використана без жодних спрощень.

Приклад 2.5. Дослідження у сфері енергоефективності «обіцяють» зменшення витрат на оплату енергоносіїв у холодну пору року в розмірі, не нижчому за 10%, у разі заміни звичайних вікон металопластиковими. Тоді усуваються «містки холоду», крізь щілини між рамою та стіною до помешкання не потрапляє холодне повітря, отже, зменшується потреба у додатковому обігріві приміщень за допомогою електроприладів та інших засобів. З метою перевірки цієї гіпотези зібрано дані про обсяги спожитої електроенергії у зимовий період мешканцями 75 квартир панельного житлового будинку в м. Києві. Виявилось, що у 25 квартирах вікна не були замінені (n_1=25), а тому середні за місяць витрати електроенергії у розрахунку на 1 м^2 загальної площі квартири сягнули **3,2 кВт•год/(міс.•м^2)** ($\tilde{x}_1 = 3{,}2$) із стандартним відхиленням (у вибірці) **S_1=0,6 кВт•год/(міс.•м^2).** У решті квартир всі вікна замінено на металопластикові (n_2=50), причому середні за місяць витрати електроенергії у розрахунку на 1 м^2 загальної площі квартири становили **2,4 кВт•год/(міс.•м^2)** ($\tilde{x}_2 = 2.4$) із стандартним відхиленням (у вибірці) **S_2=0,4 кВт•год/(міс.•м^2).** Згідно з

твердженнями фахівців математичне очікування економії електроенергії має становити:

$$\bar{x}_{1H_0} - \bar{x}_{2H_0} = \Delta\bar{x} = \frac{10\%}{100\%} \cdot 2{,}8 = 0{,}28 \text{ кВт} \cdot \text{год}/\left(\text{міс.} \cdot \text{м}^2\right).$$

Отже, H_0: $\left(\bar{x}_1 - \bar{x}_2\right)_{H_0} \le 0{,}28$, тобто H_0: «*У разі заміни звичайних вікон на металопластикові щомісячна економія електроенергії у зимовий період у розрахунку на 1 м^2* ***не перевищує 0,28 кВт•год****».*

Тоді альтернативна гіпотеза є **однобічною (правобічною)**: **H_1:** $\left(\bar{x}_1 - \bar{x}_2\right)_{H_0} > 0{,}28$ (оскільки містить знак «>»), іншими словами, *H_1: «У разі заміни звичайних вікон на металопластикові щомісячна економія електроенергії у зимовий період у розрахунку на 1 м^2* ***перевищує 0,28 кВт•год****».*

Рівень істотності встановимо в розмірі $\alpha = 0{,}05$. Хоч одна з вибірок є великою (n_2=50>30), Z-статистику для великих вибірок застосовувати не можна, адже не дотримано умови сумарного розміру обох вибірок: **n_1+n_2=25+50=75<120**. Тому слід обчислити *t*-статистику за формулою (2.5) без жодних спрощень, зважаючи на формулювання нуль-гіпотези:

$$t = \frac{\tilde{x}_1 - \tilde{x}_2 - \left(\bar{x}_{1H_0} - \bar{x}_{2H_0}\right)}{\sqrt{\frac{n_1 \cdot S_1^2(x) + n_2 \cdot S_2^2(x)}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(3{,}2 - 2{,}4) - 0{,}28}{\sqrt{\frac{25 \cdot 0{,}6^2 + 50 \cdot 0{,}4^2}{25 + 50 - 2}} \cdot \sqrt{\frac{1}{25} + \frac{1}{25}}} = \frac{0{,}52}{0{,}118} = 4{,}400.$$

Процентна точка для одностороннього критичного простору, розрахована із співвідношення $P\left(|t| \ge t^*_{\alpha;\nu=n_1+n_2-2}\right) = \alpha$ для $\alpha = 0{,}05$ становить $t^*_{0{,}05;73}$=1,666. Це число було розраховане за допомогою сучасних англомовних версій електронних таблиць Excel шляхом введення виразу **=T.INV(1-0.05;73)**. Для попередніх версій Excel до клітинки потрібно ввести функції =СТЬЮДРАСПОБР(0.05*2;73). **Підрахована статистика потрапляє до критичного простору** (як на рис. 2.3, *б*), отже, **нуль-гіпотеза виявилась хибною:** *t*=4,4>$t^*_{0{,}05;73}$=1,666. Таким чином **доведено суттєву економію електроенергії у разі заміни вікон.**

Якби були застосоввані суворіші вимоги щодо істотності верифікації, зокрема встановлено значущість на рівні $\alpha = 0{,}01$, **критична область віддалилася б**, оскільки $t^*_{0{,}01;73}$=2,379. Це число також було розраховане за допомогою Excel введенням до клітинки функції =T.INV(1-0.01;73) чи =СТЬЮДРАСПОБР(0.01*2;73) залежно від версії програми. Але й тоді **жодної підстави брати нуль-гіпотезу не виникло б**, оскільки розрахована t-статистика знов-таки не потрапила б до критичного простору t=4,4>$t^*_{0{,}01;73}$=2,379. Таким чином, **результати досліджень у сфері енергоефективності**, яку нині активно популяризують серед населення, **є істинними.** Стверджуючи це, ми можемо помилятись лише в 1% випадків.

2.3.3. Гіпотези щодо відносної частки (імовірності) (Тип «Б»)

Для верифікації гіпотез відносно частки або часток для кількох сукупностей слід досліджувати тільки **великі вибірки**, адже йдеться про імовірність, або ж параметр ***p* розкладу (закону розподілу) Бернуллі**, що, звичайно, ґрунтується на законі великих чисел. Тому в основу статистичних критеріїв верифікації таких гіпотез покладено Z-статистику. У разі істинності нуль-гіпотези розрахункова Z-статистика не потрапляє до критичного простору, процентну точку якого знайдено, виходячи з певного, обґрунтованого дослідником рівня істотності. Оскільки етапи верифікації таких гіпотез збігаються із загальною схемою (див. рис. 2.2), наведемо лише особливості формулювання нульових та альтернативних гіпотез, а також формул для розрахунку Z-статистики для випадку однієї та двох відносних часток. Звичайно для перевірки таких гіпотез потрібні, відповідно, дані однієї та двох вибірок. Обсяг кожної із вибірок має перевищувати щонайменше 20 об'єктів (***n>20*** для гіпотез щодо відносної частки одної вибірки та $n_1>20$ і $n_2>20$ для гіпотез щодо відносних часток двох вибірок).

Випадок однієї відносної частки та **однієї вибірки** розміром **n**

У такому випадку **нульова гіпотеза**, подібно до випадку гіпотези щодо середньої величини для однієї великої вибірки формулюється так: ***H₀: «*** $p = \bar{p}_{H_0}$ », тобто ***H₀: «Імовірність (відносна частота) появи досліджуваної ознаки за деякою сукупністю дорівнює певному математичному сподіванню частки»***.

Альтернативна гіпотеза залежно від **сутності проблеми** може бути**:**

- **двобічною: *H₁: «*** $p \neq \bar{p}_{H_0}$ », тобто ***H₁: «Імовірність** (відносна частота) появи досліджуваної ознаки у деякій сукупності **не дорівнює** певному математичному **сподіванню** частки»*;
- **лівобічною: *H₁: «*** $p < \bar{p}_{H_0}$ », тобто ***H₁: «Імовірність** (відносна частота) появи досліджуваної ознаки у деякій сукупності є **меншою** за певне математичне **сподівання** частки»*;
- **правобічною*: H₁: «*** $p > \bar{p}_{H_0}$ ***», тобто H₁: «Імовірність** (відносна частота) появи досліджуваної ознаки у деякій сукупності є **більшою** за певне математичне **сподівання** частки»*.

Верифікацію гіпотез виконують за допомогою Z-тесту, також подібного до випадку гіпотези щодо середньої величини для однієї великої вибірки, зважаючи на формулу дисперсії альтернативної ознаки (S=p•(1—p)/n):

$$Z = \frac{w - \bar{p}_{H_0}}{\sqrt{\frac{w \bullet (1 - w)}{n}}}, \qquad (2.6)$$

де w та $\bar{p}_{H_0}$ – значення частки (частості) відповідно у вибірці та за гіпотетичним (H_0) припущенням.

Відносну частку ***w*** визначають залежно від обсягу вибірки (***n***) та кількості спостережень у вибірці, для яких виявлено ознаку (***m***): $w = \frac{m}{n}$

Випадок двох відносних часток, а отже, і **двох вибірок** розміром **n_1** та **n_2**. При цьому вибірки можуть бути різного розміру ($n_1 \neq n_2$). У такому разі формулювання нульової та альтернативних гіпотез, а також формула обчислення Z-статистики значною мірою подібні до випадку гіпотези щодо середніх величин для двох великих вибірок:

- **нульова гіпотеза**, формулюється так: ***H_0:*** «$\bar{p}_{1H_0} = \bar{p}_{2H_0}$», або ***$H_0$:*** «$\left(\bar{p}_1 - \bar{p}_2\right)_{H_0} = 0$», а саме у вигляді висловлювання типу: ***H_0:*** ***«***Імовірність (відносна частота) появи досліджуваної ознаки у кожній з двох сукупностей є незмінною й дорівнює певному математичному сподіванню частки»;
- **альтернативна гіпотеза** залежно від **сутності проблеми** може бути**:**
  - **двобічною: *H_1:*** ***«***$\bar{p}_{1H_0} \neq \bar{p}_{2H_0}$», або ***$H_1$:*** «$\left(\bar{p}_1 - \bar{p}_2\right)_{H_0} \neq 0$», що можна сформулювати у вигляді висловлювання ***H_1:*** «Імовірність (відносна частота) появи досліджуваної ознаки по кожній із двох сукупностей **різна,** і в одній з вибірок вона **не дорівнює** певному математичному сподіванню частки, а може бути більшою чи меншою».
  - **лівобічною: *H_1: H_1: «***$\bar{p}_{1H_0} < \bar{p}_{2H_0}$», або ***$H_1$:*** «$\left(\bar{p}_1 - \bar{p}_2\right)_{H_0} < 0$», що можна сформулювати у вигляді висловлювання ***H_1: «***Імовірність (відносна частота) появи досліджуваної ознаки в одній з двох вибірок є **меншою** за певне математичне сподівання частки, що спостерігається в іншій вибірці»;
  - **правобічною: *H_1: H_1:*** «$\bar{p}_{1H_0} > \bar{p}_{2H_0}$», або ***$H_1$:*** «$\left(\bar{p}_1 - \bar{p}_2\right)_{H_0} > 0$», що можна сформулювати у вигляді висловлювання ***H_1:*** «Імовірність (відносна частота) появи досліджуваної ознаки в одній з двох вибірок **більшою** певне математичне сподівання частки, що спостерігається в іншій вибірці».

Критерієм перевірки таких гіпотез є Z-статистика, що враховує відносні частоти прояву ознаки в кожній з вибірок, відповідно $w_1 = \frac{m_1}{n_1}$ і $w_2 = \frac{m_2}{n_2}$ та усереднену частоту у всіх вибірках $\bar{p} = \frac{m_1 + m_2}{n_1 + n_2}$. З урахуванням усереднених частот Z-статистику визначають так:

$$Z = \frac{w_1 - w_2 - \left(\bar{p}_1 - \bar{p}_2\right)_{H_0}}{\sqrt{\bar{p} \bullet \left(1 - \bar{p}\right) \bullet \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{w_1 - w_2}{\sqrt{\bar{p} \bullet \left(1 - \bar{p}\right) \bullet \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad (2.7)$$

адже згідно з нуль-гіпотезою $\left(\bar{p}_1 - \bar{p}_2\right)_{H_0} = 0$,

де n_1, m_1, n_2, m_2 – відповідно обсяг та кількість спостережень, яким властива досліджувана ознака в кожній з вибірок.

Наведемо приклад верифікації гіпотез про відносні частки (імовірності) однієї вибірки.

> **Приклад 2.6.** Чи правда, що менш ніж 15% покупців користуються купонами для придбання акційних товарів зі знижками ($\bar{p}_{H_0} = 0{,}15 = 15\%$), які зазвичай друкують на останніх сторінках рекламних буклетів, якщо, спостерігаючи за розрахунками біля кас зафіксовано, що з **64 покупців (n=64)** акційних товарів купон на знижку пред'явило лише **6 осіб (m=6).** Для перевірки гіпотези взято рівень значущості **5%.**

Нульову гіпотезу сформулюємо, виходячи з припущення про відсутність значних розбіжностей між гіпотетичною та фактичною частотою використання купонів на знижку:

H_0: « $p \geq 15\%$ ***»,*** тобто *H_0: «Відносна частота використання купонів на придбання зі знижками акційних товарів є більшою або дорівнює 15%».*

Тоді альтернативна гіпотеза є **лівобічною** (оскільки містить знак «<») ***H_1: «*** $p < 15\%$ », тобто ***H_1: «****Імовірність використання купонів на придбання зі знижками акційних товарів є меншою, ніж 15%*».

Відповідно до даних вибірки частка покупок з використанням купонів дорівнює $w = \frac{m}{n} = \frac{6}{64} = 0{,}09375$. Тоді Z-статистика за формулою (2.6) становитиме:

$$Z = \frac{w - \bar{p}_{H_0}}{\sqrt{\frac{w \cdot (1 - w)}{n}}} = \frac{0{,}09375 - 0{,}15}{\sqrt{\frac{0{,}09375 \cdot (1 - 0{,}09375)}{64}}} \approx -1{,}544 .$$

Згідно з табл. 2.5 для лівобічної критичної області з рівнем істотності $\alpha = 0{,}05$ критичне значення $Z^*_{0{,}05} = -1{,}64$. Отже, розрахована Z-статистика не потрапляє до критичної області:

$$\mathbf{|Z|=1{,}55<|Z^*_{0{,}05}|=1{,}64}.$$

Таким чином, підстав для відхилення нуль-гіпотези немає. Отже, на рівні істотності $\alpha = 0{,}05$ імовірність використання купонів на придбання акційних товарів зі знижками, надрукованих на останніх сторінках рекламних буклетів, є не меншою, ніж 15%. Помилковим цей висновок може бути лише в п'ятьох випадках зі 100.

У наступному прикладі розкрито процедуру верифікації гіпотез про відносні частки для двох вибірок.

Приклад 2.7. Менеджери мережі магазинів з продажу будівельних матеріалів вирішили з'ясувати, як робота аніматорів-промоутерів впливає на продаж емалі. В одному магазині систематично протягом місяця працювали аніматори, звертаючи увагу відвідувачів на емалі незалежно від мети відвідування магазину. При цьому з 1600 осіб (n_1=1600), що заходили до магазину, покупки зробили 280 (m_1=280). В іншому магазині, де промоутерів не було, зі **1300** (n_2=1300) відвідувачів пішли з магазину з покупками 192 особи (m_2=192). З'ясуємо позитивний вплив промоакцій на рішення придбати будівельну емаль із рівнем значущості 0,05 та 0,01(тобто $\alpha_1 = 0{,}05$, $\alpha_2 = 0{,}01$).

Нульову гіпотезу сформулюємо, виходячи із припущення про не значні розбіжності між частотою покупок, зумовлених роботою аніматорів у тому чи іншому магазині, зокрема і збільшенням частоти покупок завдяки роботі промо-персоналу, тобто

H₀: «$\bar{p}_{1H_0} = \bar{p}_{2H_0}$», або **H₀:** «$\left(\bar{p}_1 - \bar{p}_2\right)_{H_0} = 0$».

Тоді альтернативна гіпотеза є **правобічною** (оскільки містить знак «>») ***H₁*:**

H₁: «$\bar{p}_{1H_0} > \bar{p}_{2H_0}$», або ***H₁:*** «$\left(\bar{p}_1 - \bar{p}_2\right)_{H_0} > 0$»,

тобто

H₁: «*Імовірність покупки є вищою у тих магазинах, де працюють аніматори-промоутери*».

Відносна частка покупців, що придбали фарбу-емаль в магазині, де працювали промоутери: $w_1 = \frac{m_1}{n_1} = \frac{280}{1600} = 0{,}175$, або 17,5%.

Відносна частка покупців, що придбали фарбу-емаль в магазині, де не працювали промоутери: $w_2 = \frac{m_2}{n_2} = \frac{192}{1300} = 0{,}148$, або 14,8%.

Відносна частка покупок за обома вибірками: $\bar{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{280 + 192}{1600 + 1300} = \frac{742}{2900} = 0{,}163$, або 16,3%.

Тоді Z-статистика (2.7) становитиме:

$$Z = \frac{w_1 - w_2}{\sqrt{\bar{p} \cdot \left(1 - \bar{p}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0{,}175 - 0{,}148}{\sqrt{0{,}163 \cdot \left(1 - 0{,}163\right) \cdot \left(\frac{1}{1600} + \frac{1}{1300}\right)}} = 1{,}981.$$

Відповідно до табл. 2.5 для правобічної критичної області з рівнем істотності $\alpha_1 = 0{,}05$ критичне значення $Z^*_{0{,}05}$ =1,64. Отже, розрахована Z-статистика потрапляє до критичної області:

$Z=1{,}981 > Z^*_{0{,}05}=1{,}64$.

Таким чином, нуль-гіпотезу про брак суттєвого ефекту від роботи промоутерів слід відхилити. Отже, на рівні істотності $\alpha_1 = 0{,}05$ робота промоутерів-аніматорів у магазинах будівельних матеріалів суттєво впливає на частоту покупок емалі. Помилковим цей висновок може бути лише в п'ятьох випадках зі 100.

Однак, обмежуючи рівень імовірності помилки 1%, тобто для $\alpha_2 = 0{,}01$ за даними табл. 2.5, для правобічної критичної області слід обрати **критичне значення $Z^*_{0,01}$=2,33**. Тоді **розрахована Z-статистика не потрапить до критичної області:**

$$Z=1{,}981<Z^*_{0,01}=2{,}33.$$

Таким чином, зникають підстави для відхилення нуль-гіпотези про брак суттєвого ефекту від роботи промоутерів. Тобто на рівні істотності $\alpha_2 = 0{,}01$ робота промоутерів-аніматорів у магазинах будівельних матеріалів суттєво не впливає на частоту покупок емалі, і цей висновок може бути хибним лише в одному випадку зі ста. Отже, посилення вимог щодо значущості статистичної гіпотези суттєво вплинуло на розв'язок верифікації.

2.3.4. ГІПОТЕЗИ ЩОДО ДИСПЕРСІЇ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ (ТИП «В»)

Випадок однієї сукупності. Мала вибірка, n<30

Дисперсія популяції є важливим параметром генеральної сукупності, особливо тоді, коли одиниці такої сукупності не повинні відрізнятись одна від одної. Зокрема, це стосується перевірки якості деталей, вузлів чи з'єднань в процесі виробництва продукції або точності роботи вимірювальних приладів, адже у такому випадку мірою якості вимірювання слугує дисперсія, або стандартне відхилення.

Зазвичай висновки про дисперсію доводиться робити на підставі малої вибірки, оскільки неможливо піддавати суцільному контролю усю партію виробів, а надто тоді, коли контрольні вимірювання спричинюють руйнування зразків продукції чи спрацювання виробів до цілковитої відмови. Нульова гіпотеза тоді має такий вигляд: ***H_0:*** «$\sigma^2 \le \sigma^2_{H_0}$», тобто ***$H_0$:*** «Дисперсія в деякій сукупності дорівнює (не перевищує) певного математичного сподівання дисперсії ($\sigma^2_{H_0}$)». **Альтернативна гіпотеза,** з огляду на сутність проблеми, **завжди є правобічною: *H_1: «***$\sigma^2 > \sigma^2_{H_0}$», тобто завжди містить знак «більше» (>), та формулюють її так ***H_1: «***Дисперсія у деякій сукупності є більшою за певне математичне сподівання дисперсії ($\sigma^2_{H_0}$)». Верифікацію таких гіпотез виконують за допомогою статистики Хі-квадрат (χ^2):

$$\chi^2 = \frac{n \bullet S^2(x)}{\sigma^2_{H_0}}, \qquad (2.8)$$

де $\sigma^2_{H_0}$, $S^2(x)$ – дисперсія за гіпотетичним (H_0) припущенням та розрахована за даними вибірки;

n – обсяг вибірки.

Якщо значення тесту **перевищить табличне (критичне)** значення **статистики розподілу** Пірсона Хі-квадрат ($\left(\chi^2_{\alpha;\nu=n-1}\right)^*$) за обраного дослідником рівня істотності α і кількості ступенів вільності $\nu = n-1$ (табл. 2.7), від якої залежить цей розподіл, нульову гіпотезу про відсутність перевищення дисперсією генеральної сукупності очікуваному значенню $\sigma^2_{H_0}$ слід відкинути та обрати альтернативну:

$$\chi^2 > \left(\chi^2_{\alpha;\nu=n-1}\right)^* \Rightarrow \text{H}_0 \Rightarrow \text{H}_1(\text{i}).$$

Таблиця 2.7

Критичні значення $\left(\chi^2_{\alpha;\nu=n-1}\right)^*$

ν	**Двобічна критична область**					
	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}02$	$\alpha = 0{,}01$	$\alpha = 0{,}005$
1	2,706	3,841	5,024	5,412	6,635	7,879
2	4,605	5,991	7,378	7,824	9,210	10,597
3	6,251	7,815	9,348	9,837	11,345	12,838
4	7,779	9,488	11,143	11,668	13,277	14,860
5	9,236	11,070	12,833	13,388	15,086	16,750
6	10,645	12,592	14,449	15,033	16,812	18,548
7	12,017	14,067	16,013	16,622	18,475	20,278
8	13,362	15,507	17,535	18,168	20,090	21,955
9	14,684	16,919	19,023	19,679	21,666	23,589
10	15,987	18,307	20,483	21,161	23,209	25,188

Закінчення табл. 2.7

ν	**Двобічна критична область**					
	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}02$	$\alpha = 0{,}01$	$\alpha = 0{,}005$
11	17,275	19,675	21,920	22,618	24,725	26,757
12	18,549	21,026	23,337	24,054	26,217	28,300
13	19,812	22,362	24,736	25,472	27,688	29,819
14	21,064	23,685	26,119	26,873	29,141	31,319
15	22,307	24,996	27,488	28,259	30,578	32,801
16	23,542	26,296	28,845	29,633	32,000	34,267
17	24,769	27,587	30,191	30,995	33,409	35,718
18	25,989	28,869	31,526	32,346	34,805	37,156
19	27,204	30,144	32,852	33,687	36,191	38,582
20	28,412	31,410	34,170	35,020	37,566	39,997
21	29,615	32,671	35,479	36,343	38,932	41,401
22	30,813	33,924	36,781	37,659	40,289	42,796
23	32,007	35,172	38,076	38,968	41,638	44,181
24	33,196	36,415	39,364	40,270	42,980	45,559

ν	**Двобічна критична область**					
	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}02$	$\alpha = 0{,}01$	$\alpha = 0{,}005$
25	34,382	37,652	40,646	41,566	44,314	46,928
26	35,563	38,885	41,923	42,856	45,642	48,290
27	36,741	40,113	43,195	44,140	46,963	49,645
28	37,916	41,337	44,461	45,419	48,278	50,993
29	39,087	42,557	45,722	46,693	49,588	52,336
30	40,256	43,773	46,979	47,962	50,892	53,672

Тобто в задачах про перевірку якості слід дійти висновку, що якість продукції, деталей, робіт не задовольняє стандарту, оскільки розмах варіації відхилень перевищує регламентовані допуски. При цьому залишається імовірність ухвалення помилкового рішення на рівні істотності α.

В іншому випадку, коли значення статистики Хі-квадрат виявиться меншим, ніж табличне, за рівня істотності α та кількості ступенів вільності на одиницю меншою за кількість спостережень**,** нульову гіпотезу про відповідність партії продукції стандарту варто взяти: $\chi^2 \le \left(\chi^2_{\alpha;\nu=n-1}\right)^* \Rightarrow \mathrm{H}_0$ (і), адже тоді дисперсія в деякій сукупності (усій партії виробленої продукції) не перевищує певного математичного сподівання дисперсії ($\sigma^2{}_{H_0}$), принаймні у (1—α)•100% випадків.

Приклад 2.8. Граничні відхилення лінійних розмірів конструктивних сталевих елементів та блоків розміром понад 4,0 до 8 м становить $\pm$ 6мм. Згідно з правилом шістьох сигм (максимальний розмір варіації – три стандартні відхилення) одне стандартне відхилення не повинне перевищувати 2 мм: $\sigma \le 2$ мм. Отже, відповідно до стандарту[4] максимальна допустима дисперсія металоконструкцій, що надходять на монтаж, становить $\sigma^2{}_{H_0} = 4$ мм^2. На склад будівельного підприємства надійшло 150 сталевих арматурних стержнів періодичного профілю завдовжки 6 м. Перевірено вибірку – 15 стержнів (n=15), за даними якої встановлено розміри арматури (табл. 2.8). В цій таблиці наведено також розрахункові показники, потрібні для обчислення дисперсії у вибірці: середня довжина стержня $\bar{l} = 5999{,}07$ мм**,** середній квадрат довжини стержня $\overline{l^2} = 35988808$ мм^2. Слід обґрунтувати висновок про якість отриманої партії арматури, її придатність для монтажу будівельних конструкцій та доцільність оформлення претензії до заводу виробника із подальшою заміною неякісної продукції.

Таблиця 2.8

Результати обстеження арматурних стержнів

[4] СНиП III-18-75 «Правила виробництва і приймання робіт. Металеві конструкції».

Номер стержня							
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Довжина стержня, l, мм, середнє значення – 5999,07 мм							
6002	6003	5997	5998	6001	5994	5995	5999
Розрахунковий показник – квадрат довжини стержня, l^2, мм^2, середнє значення – 35988808 мм^2							
36024004	36036009	35964009	35976004	36012001	35928036	35940025	35988001
Номер стержня							
9	**10**	**11**	**12**	**13**	**14**	**15**	
Довжина стержня, l, мм, середнє значення– 5999,07 мм							
5999	6001	5999	6001	5997	6003	5997	
Розрахунковий показник – квадрат довжини стержня, l^2, мм^2, середнє значення – 35988808 мм^2							
35988001	36012001	35988001	36012001	35964009	36036009	35964009	

Нульову гіпотезу формульовано так: ***H₀***: «$\sigma^2 \le 4$ мм^2», тобто ***H₀:*** «Дисперсія отриманої партії сталевої арматури не перевищує 4 мм^2». **Альтернативна гіпотеза,** яка **завжди є правобічною і завжди містить знак «більше» (>),** має такий вигляд**: H_1: «**$\sigma^2 > 4$ мм^2». Альтернативна гіпотеза має таке формулювання ***H₁***: «Дисперсія отриманої партії арматури є **більшою за** 4 мм^2». Зважаючи на суспільно-економічну значущість безпеки будівлі, беремо рівень істотності $\alpha = 0{,}005$ **.**

Для того щоби здійснити верифікацію гіпотези, запобігаючи використанню неякісних конструкцій у будівництві, спочатку потрібно розрахувати дисперсію довжини стержня в обстеженій вибірці. За даними табл. 2.8 дисперсія довжини стержня у вибірці дорівнюватиме:

$$S^2(l) = \overline{l^2} - \bar{l}^2 = 35988808 - 5999{,}07^2 = 7{,}13 \text{ мм}^2.$$

Тоді статистика Хі-квадрат (χ^2) становитиме:

$$\chi^2 = \frac{n \cdot S^2(x)}{\sigma^2_{H_0}} = \frac{15 \cdot 7{,}13}{4} = 26{,}73.$$

Згідно з табл. 2.7, критичне значення статистики розподілу Пірсона Хі-квадрат ($\left(\chi^2_{0{,}005;14}\right)^* = 31{,}319$) за обраного дослідником рівня істотності $\alpha = 0{,}005$ і кількості ступенів вільності $\nu = n - 1 = 15 - = 14$. Цю кількість можна розрахувати за допомогою Excel, увівши до клітинки функцій =CHISQ.INV.RT(0.005;14) або =ХИ2ОБР(0,005;14) залежно від версії програми.

Статистика Хі-квадрат у вибірці не перевищила критичного значення: $\chi^2 = 26{,}73 \le \left(\chi^2_{\alpha;\nu=n-1}\right)^* = 31{,}319 \Rightarrow H_0$ (і), тобто дисперсія в отриманій партії арматурних стержнів не перевищує дисперсії, дозволеної стандартом ($\sigma^2_{H_0} = 4$ мм^2) у 99,5%(=(1—0,005)•100%) випадків. Тобто провадити

претензійну роботу із заводом-виробником та вимагати заміни стержнів недоцільно.

Випадок двох сукупностей. Мала вибірка, n<30

Дослідження рівності дисперсій двох сукупностей дає змогу порівнювати міру розсіювання певної ознаки цих двох популяцій. Формулювання та верифікація гіпотез потребують нормального розподілу досліджуваної ознаки в обох вибірках. За розрахунками дисперсій щодо кожної із вибірок (які зазвичай не перевищують 30 одиниць, а отже, належать до малих, причому розмір вибірок може бути різним ($n_1 \neq n_2$)), формулюють **нульову гіпотезу,** що, подібно до порівняння середнього у двох вибірках, також може мати два варіанти формулювання: H_0: $\sigma_1^2 = \sigma_2^2$ або ж H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$, якщо йдеться про рівність дисперсій у двох популяціях.

Як й у випадку однієї вибірки, альтернативна гіпотеза завжди є правобічною і завжди містить знак «більше» (>): H_1: «$\sigma_1^2 > \sigma_2^2$», або H_1: $\frac{\sigma_1^2}{\sigma_2^2} > 1$, адже вона означає перевищення дисперсією однієї з вибірок дисперсії іншої. Тест істотності, за допомогою якого ухвалюється рішення про верифікацію (2.9), ґрунтується на розподілі Фішера–Снедекора із кількістю ступенів вільности, що стоять у знаменнику ($\nu_1 = n_1 - 1$) та чисельнику ($\nu_2 = n_2 - 1$) формули (2.9):

$$S = \frac{\frac{n_1 \bullet S_1^2(x)}{n_1 - 1}}{\frac{n_2 \bullet S_2^2(x)}{n_2 - 1}} = \frac{n_1 \bullet S_1^2(x)}{n_1 - 1} \bullet \frac{n_2 - 1}{n_2 \bullet S_2^2(x)}, \qquad (2.9)$$

де $S_1^2(x)$, $S_2^2(x)$ – дисперсія відповідно першої та другої вибірок. **Номер «1» одержує та вибірка**, **у якій дисперсія більша**. Вибірку з меншою дисперсією, незалежно від розміру та значень ознаки, нумерують індексом «2» відповідно до вимог альтернативної гіпотези;

n_1, n_2 – відповідно обсяг вибірки з більшою (1) та меншою (2) дисперсіями.

Найчастіше рівень істотності під час перевірки таких гіпотез становить $\alpha = 0{,}05$. Для цього випадку критичні значення F-статистики за кількості ступенів вільності ν_1 і ν_2 наведено в табл. 2.9. Втім, для будь-якого рівня значущості та кількості ступенів вільності можна одержати табличне значення F-критерію за допомогою функції Excel FРАСПОБР з категорії статистичних, синтаксис якої потребує введення в дужках показника істотності у вигляді десяткового дробу, а також кількості ступенів вільності із знаменника та чисельника. Кожен з цих трьох аргументів відокремлюється один від одного крапкою з комою (;).

Нульову гіпотезу відкидають і беруть альтернативну у разі перевищення F-статистикою, обчисленою за вибірками, критичного значення $F_{\alpha;\nu_1=n_1-1;\nu_2=n_2-1}$* (або $F_{\alpha;\nu_1;\nu_2}$*):

$$F > F_{\alpha;\nu_1;\nu_2}^* \Rightarrow \text{~~H}_0\text{~~} \Rightarrow \text{H}_1\text{(i)}.$$

Наведемо приклад верифікації гіпотези про рівність дисперсій двох вибірок, використавши дані прикладу верифікації гіпотези про споживання електроенергії мешканцями квартир з різним типом віконних конструкцій.

Приклад 2.9. У цьому прикладі вибірка з номером «1» характеризувалась більшою мінливістю спожитих кіловат-годин електроенергії у розрахунку на 1 м^2 загальної площі житла, тож індексацію умовних позначень змінювати не потрібно:

- для підвибірки квартир, у яких не замінили вікна на металопластикові: n_1=25, **S_1=0,6 кВт•год/(міс.•м^2).** Ця підвибірка має номер «1», хоча її розмір менший**;**
- для підвибірки квартир, у яких замінено вікна на металопластикові: n_2=50, **S_2=0,4 кВт•год/(міс.•м^2).** Ця підвибірка має номер «2», незважаючи на її більшу чисельність, потреба в електроенергії виявилась більш однорідною**.**

Таблиця 2.9

F-розподіл (Фішера — Снедекора)

(в чисельнику значення для $\alpha = 0{,}05;\ P = 0{,}95$ в знаменнику – для $\alpha = 0{,}01;\ P = 0{,}99$)

Кількість ступенів вільності		n_2 — показник з чисельника																						
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**25**	**50**	**100**
n_1 — показник із знаменника	1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882	242.983	243.906	244.690	245.364	245.950	246.464	246.918	247.323	247.686	248.013	249.260	251.774	253.041
		4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847	6083.317	6106.321	6125.865	6142.674	6157.285	6170.101	6181.435	6191.529	6200.576	6208.730	6239.825	6302.517	6334.110
	2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.405	19.413	19.419	19.424	19.429	19.433	19.437	19.440	19.443	19.446	19.456	19.476	19.486
		98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.408	99.416	99.422	99.428	99.433	99.437	99.440	99.444	99.447	99.449	99.459	99.479	99.489
	3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.763	8.745	8.729	8.715	8.703	8.692	8.683	8.675	8.667	8.660	8.634	8.581	8.554
		34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.133	27.052	26.983	26.924	26.872	26.827	26.787	26.751	26.719	26.690	26.579	26.354	26.240
	4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.936	5.912	5.891	5.873	5.858	5.844	5.832	5.821	5.811	5.803	5.769	5.699	5.664
		21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.452	14.374	14.307	14.249	14.198	14.154	14.115	14.080	14.048	14.020	13.911	13.690	13.577
	5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.704	4.678	4.655	4.636	4.619	4.604	4.590	4.579	4.568	4.558	4.521	4.444	4.405
		16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.963	9.888	9.825	9.770	9.722	9.680	9.643	9.610	9.580	9.553	9.449	9.238	9.130
	6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.027	4.000	3.976	3.956	3.938	3.922	3.908	3.896	3.884	3.874	3.835	3.754	3.712
		13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.790	7.718	7.657	7.605	7.559	7.519	7.483	7.451	7.422	7.396	7.296	7.091	6.987
	7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.603	3.575	3.550	3.529	3.511	3.494	3.480	3.467	3.455	3.445	3.404	3.319	3.275
		12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.538	6.469	6.410	6.359	6.314	6.275	6.240	6.209	6.181	6.155	6.058	5.858	5.755
	8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	3.313	3.284	3.259	3.237	3.218	3.202	3.187	3.173	3.161	3.150	3.108	3.020	2.975
		11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.734	5.667	5.609	5.559	5.515	5.477	5.442	5.412	5.384	5.359	5.263	5.065	4.963
	9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.102	3.073	3.048	3.025	3.006	2.989	2.974	2.960	2.948	2.936	2.893	2.803	2.756
		10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.178	5.111	5.055	5.005	4.962	4.924	4.890	4.860	4.833	4.808	4.713	4.517	4.415
	10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.943	2.913	2.887	2.865	2.845	2.828	2.812	2.798	2.785	2.774	2.730	2.637	2.588
		10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.772	4.706	4.650	4.601	4.558	4.520	4.487	4.457	4.430	4.405	4.311	4.115	4.014
	11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.818	2.788	2.761	2.739	2.719	2.701	2.685	2.671	2.658	2.646	2.601	2.507	2.457
		9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.462	4.397	4.342	4.293	4.251	4.213	4.180	4.150	4.123	4.099	4.005	3.810	3.708

Продовження табл. 2.9

Кількість ступенів вільності		n_2 — показник з чисельника																						
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**25**	**50**	**100**
n_1 — показник із знаменника	12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.717	2.687	2.660	2.637	2.617	2.599	2.583	2.568	2.555	2.544	2.498	2.401	2.350
		9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.220	4.155	4.100	4.052	4.010	3.972	3.939	3.909	3.883	3.858	3.765	3.569	3.467
	13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.635	2.604	2.577	2.554	2.533	2.515	2.499	2.484	2.471	2.459	2.412	2.314	2.261
		9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	4.025	3.960	3.905	3.857	3.815	3.778	3.745	3.716	3.689	3.665	3.571	3.375	3.272
	14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.565	2.534	2.507	2.484	2.463	2.445	2.428	2.413	2.400	2.388	2.341	2.241	2.187
		8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.864	3.800	3.745	3.698	3.656	3.619	3.586	3.556	3.529	3.505	3.412	3.215	3.112
	15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.507	2.475	2.448	2.424	2.403	2.385	2.368	2.353	2.340	2.328	2.280	2.178	2.123
		8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.730	3.666	3.612	3.564	3.522	3.485	3.452	3.423	3.396	3.372	3.278	3.081	2.977
	16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.456	2.425	2.397	2.373	2.352	2.333	2.317	2.302	2.288	2.276	2.227	2.124	2.068
		8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.616	3.553	3.498	3.451	3.409	3.372	3.339	3.310	3.283	3.259	3.165	2.967	2.863
	17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.413	2.381	2.353	2.329	2.308	2.289	2.272	2.257	2.243	2.230	2.181	2.077	2.020
		8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.519	3.455	3.401	3.353	3.312	3.275	3.242	3.212	3.186	3.162	3.068	2.869	2.764
	18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.374	2.342	2.314	2.290	2.269	2.250	2.233	2.217	2.203	2.191	2.141	2.035	1.978
		8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.434	3.371	3.316	3.269	3.227	3.190	3.158	3.128	3.101	3.077	2.983	2.784	2.678
	19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.340	2.308	2.280	2.256	2.234	2.215	2.198	2.182	2.168	2.155	2.106	1.999	1.940
		8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.360	3.297	3.242	3.195	3.153	3.116	3.084	3.054	3.027	3.003	2.909	2.709	2.602
	20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.310	2.278	2.250	2.225	2.203	2.184	2.167	2.151	2.137	2.124	2.074	1.966	1.907
		8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.294	3.231	3.177	3.130	3.088	3.051	3.018	2.989	2.962	2.938	2.843	2.643	2.535
	21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.283	2.250	2.222	2.197	2.176	2.156	2.139	2.123	2.109	2.096	2.045	1.936	1.876
		8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.236	3.173	3.119	3.072	3.030	2.993	2.960	2.931	2.904	2.880	2.785	2.584	2.475
	22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.259	2.226	2.198	2.173	2.151	2.131	2.114	2.098	2.084	2.071	2.020	1.909	1.849
		7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.184	3.121	3.067	3.019	2.978	2.941	2.908	2.879	2.852	2.827	2.733	2.531	2.422
	23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.236	2.204	2.175	2.150	2.128	2.109	2.091	2.075	2.061	2.048	1.996	1.885	1.823
		7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.137	3.074	3.020	2.973	2.931	2.894	2.861	2.832	2.805	2.781	2.686	2.483	2.373

Закінчення табл. 2.9

Кількість ступенів вільності		n_2 — показник з чисельника																						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	50	100
n_1 — показник із знаменника	24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.216	2.183	2.155	2.130	2.108	2.088	2.070	2.054	2.040	2.027	1.975	1.863	1.800
		7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.094	3.032	2.977	2.930	2.889	2.852	2.819	2.789	2.762	2.738	2.643	2.440	2.329
	25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.198	2.165	2.136	2.111	2.089	2.069	2.051	2.035	2.021	2.007	1.955	1.842	1.779
		7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	3.056	2.993	2.939	2.892	2.850	2.813	2.780	2.751	2.724	2.699	2.604	2.400	2.289
	30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.126	2.092	2.063	2.037	2.015	1.995	1.976	1.960	1.945	1.932	1.878	1.761	1.695
		7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.906	2.843	2.789	2.742	2.700	2.663	2.630	2.600	2.573	2.549	2.453	2.245	2.131
	40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.038	2.003	1.974	1.948	1.924	1.904	1.885	1.868	1.853	1.839	1.783	1.660	1.589
		7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.727	2.665	2.611	2.563	2.522	2.484	2.451	2.421	2.394	2.369	2.271	2.058	1.938
	50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026	1.986	1.952	1.921	1.895	1.871	1.850	1.831	1.814	1.798	1.784	1.727	1.599	1.525
		7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698	2.625	2.562	2.508	2.461	2.419	2.382	2.348	2.318	2.290	2.265	2.167	1.949	1.825
	60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.952	1.917	1.887	1.860	1.836	1.815	1.796	1.778	1.763	1.748	1.690	1.559	1.481
		7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.559	2.496	2.442	2.394	2.352	2.315	2.281	2.251	2.223	2.198	2.098	1.877	1.749
	70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969	1.928	1.893	1.863	1.836	1.812	1.790	1.771	1.753	1.737	1.722	1.664	1.530	1.450
		7.011	4.922	4.074	3.600	3.291	3.071	2.906	2.777	2.672	2.585	2.512	2.450	2.395	2.348	2.306	2.268	2.234	2.204	2.176	2.150	2.050	1.826	1.695
	80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951	1.910	1.875	1.845	1.817	1.793	1.772	1.752	1.734	1.718	1.703	1.644	1.508	1.426
		6.963	4.881	4.036	3.563	3.255	3.036	2.871	2.742	2.637	2.551	2.478	2.415	2.361	2.313	2.271	2.233	2.199	2.169	2.141	2.115	2.015	1.788	1.655
	90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938	1.897	1.861	1.830	1.803	1.779	1.757	1.737	1.720	1.703	1.688	1.629	1.491	1.407
		6.925	4.849	4.007	3.535	3.228	3.009	2.845	2.715	2.611	2.524	2.451	2.389	2.334	2.286	2.244	2.206	2.172	2.142	2.114	2.088	1.987	1.759	1.623
	100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927	1.886	1.850	1.819	1.792	1.768	1.746	1.726	1.708	1.691	1.676	1.616	1.477	1.392
		6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503	2.430	2.368	2.313	2.265	2.223	2.185	2.151	2.120	2.092	2.067	1.965	1.735	1.598
	120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910	1.869	1.834	1.803	1.775	1.750	1.728	1.709	1.690	1.674	1.659	1.598	1.457	1.369
		6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472	2.399	2.336	2.282	2.234	2.192	2.154	2.119	2.089	2.060	2.035	1.932	1.700	1.559
	150	3.904	3.056	2.665	2.432	2.274	2.160	2.071	2.001	1.943	1.894	1.853	1.817	1.786	1.758	1.734	1.711	1.691	1.673	1.656	1.641	1.580	1.436	1.345
		6.807	4.749	3.915	3.447	3.142	2.924	2.761	2.632	2.528	2.441	2.368	2.305	2.251	2.203	2.160	2.122	2.088	2.057	2.029	2.003	1.900	1.665	1.520
	200	3.888	3.041	2.650	2.417	2.259	2.144	2.056	1.985	1.927	1.878	1.837	1.801	1.769	1.742	1.717	1.694	1.674	1.656	1.639	1.623	1.561	1.415	1.321
		6.763	4.713	3.881	3.414	3.110	2.893	2.730	2.601	2.497	2.411	2.338	2.275	2.220	2.172	2.129	2.091	2.057	2.026	1.997	1.971	1.868	1.629	1.481

Отже, H_0: $\sigma_1^2 = \sigma_2^2$, тобто *H_0:* «Мінливість потреби в електроенергії у зимовий період не залежить від того, чи встановлено у помешканні металопластикові вікна, чи ні». Альтернативна гіпотеза, як уже зазначено, є **правобічною**: ***H_1:*** $\sigma_1^2 > \sigma_2^2$ (оскільки містить знак «>»), іншими словами, *H_1: «Без заміни звичайних вікон металопластиковими мінливість (диференціація) потреби в електроенергії у зимовий період зростає».* Рівень істотності встановимо в розмірі $\alpha = 0{,}05$. Тоді F-статистика за формулою (2.9) становитиме:

$$S = \frac{n_1 \bullet S_1^2(x)}{n_1 - 1} \bullet \frac{n_2 - 1}{n_2 \bullet S_2^2(x)} = \frac{25 \bullet 0{,}6^2}{25 - 1} \bullet \frac{50 - 1}{50 \bullet 0{,}4^2} = \frac{9 \bullet 49}{24 \bullet 8} = 2{,}297 .$$

Критичне значення F-критерію для $\alpha = 0{,}05$ та кількості ступенів вільності $\nu_1 = n_1 - 1 = 25 - 1 = 24$ (у знаменнику) та $\nu_2 = n_2 - 1 = 50 - 1 = 49$ (у чисельнику) становить $F^*_{0{,}05;24;49} = 1{,}742$. Це число було розраховане за допомогою англомовної версії Excel введенням до клітинки функції =F.INV.RT(0.05;24;49). У назві цієї функції є символи RT – це скорочене англійське «right tail», тобто «правий хвіст», що узгоджується з типом гіпотези – правобічна. У старих версіях табличного процесору використовують функцію = FРАСПОБР(0,05;24;49).

Підрахована статистика потрапляє до критичного простору (як на рис. 2.3, *б*), отже, нуль-гіпотеза виявилась хибною: $F=2{,}297 > F^*_{0{,}05;24;49}=1{,}742$. Таким чином, доведено суттєвість зменшення варіації потреби в електроенергії у разі заміни вікон.

Якби були взяті суворіші вимоги щодо істотності верифікації, зокрема встановлено значущість на рівні $\alpha = 0{,}01$, критична область віддалилася б, оскільки $F^*_{0{,}05;24;49} = 2{,}192$. Це число також було розраховане за допомогою Excel введенням до клітинки функцій =F.INV.RT(0.01;24;49), =СТЬЮДРАСПОБР(0,01;24;49), відповідно до версії табличного процесора. Але й тоді жодної підстави приймати нуль-гіпотезу не виникло б, оскільки розрахована F-статистика знов-таки не потрапила б до критичного простору $F = 2{,}297 > F^*_{0{,}051;24;49} = 2{,}192$. Таким чином, заходи з підвищення енергоефективності не лише зменшують, а й стабілізують потреби мешканців в електричній енергії в холодну пору року. Стверджуючи це, ми можемо помилятись лише в 1% випадків.

Три і більше сукупностей. Велика або мала вибірка, розподілена на декілька підгруп

У таких випадках виконують дисперсійний аналіз, що дає змогу порівняти мінливість певної ознаки між групами та усередині груп. Міжгрупова дисперсія вказує на суттєві розбіжності середніх групових величин, а дисперсія всередині груп, відома як залишкова дисперсія, характеризує випадкову помилку розподілу спостережень за групами, виконаного на основі

досліджуваної ознаки. Цей вид аналізу має також назву «аналіз дисперсії в одиничній класифікації». Якщо поділ на групи (класифікацію) виконують, зважаючи на різний рівень прояву досліджуваної ознаки внаслідок дії певного фактора, то дисперсійний аналіз дає змогу виявити вплив інтенсивності дії такого фактора на рівень прояву досліджуваної властивості.

Дослідження рівності дисперсій декількох (m) сукупностей виконують, виходячи з припущення, що розподіл ознаки в усіх сукупностях є близьким до нормального. За таких умов вважають, що дисперсії усіх популяцій рівні, тому перший етап верифікації гіпотез, як в усіх уже розглянутих випадках, починається з формулювання нульової гіпотези стосовно рівності середнього рівня ознаки у популяціях:

$$H_0: \bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_m. \quad (2.10)$$

Тоді альтернативна гіпотеза спростовуватиме нульову:

$$H_1: \textit{не правильно, що } \bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_m. \quad (2.11)$$

Для перевірки нульової гіпотези з кожної популяції обирають незалежні вибірки, приблизно однакового розміру (n_j, $\sum_{j=1}^{m} n_j = n$), на основі яких обчислюють тест істотності, відомий як статистика Фішера – Снедекора. Для обчислення цього тесту спочатку виконують низку підготовчих обчислень за даними значень ознаки елементів вибірок, x_{ij} ($i = \overline{1, n}$; $j = \overline{1, m}$), причому індекс ***i*** характеризує порядковий номер елемента у вибірці, а індекс ***j*** визначає порядковий номер вибірки (сукупності). Для розрахунку статистики Фішера – Снедекора насамперед потрібні значення загальної середньої у всіх вибірках ($\bar{x}$), а також середніх групових величин ($\bar{x}_j$) відповідно до формул:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}, \quad (2.12)$$

$$\bar{x} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{ij}}{\sum_{j=1}^{m} n_j} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{ij}}{n}. \quad (2.13)$$

На основі обчислених за формулами (2.12), (2.13) середніх розраховують такі дисперсійні суми квадратів:

– суму, що відображає варіацію між групам и $\sum_{j=1}^{m}\left(\overline{x_i}-\overline{x}\right)^2\cdot n_j$, причому кількість ступенів вільності на одну менша за кількість груп (*m—1*);

– суму, яка визначає варіації всередині груп, тобто залишкову варіацію $\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x_j}\right)^2$ з кількістю ступенів вільності, що дорівнює різниці між загальним обсягом всіх вибірок та кількістю вибірок (*n—m*).

Статистику Фішера – Снедекора визначають якраз з урахуванням наведених дисперсійних сум квадратів та показників кількості ступенів вільності:

$$F=\frac{\dfrac{\sum_{j=1}^{m}\left(\overline{x_i}-\overline{x}\right)^2\cdot n_j}{m-1}}{\dfrac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x_j}\right)^2}{n-m}}=\frac{\sum_{j=1}^{m}\left(\overline{x_i}-\overline{x}\right)^2\cdot n_j}{m-1}\cdot\frac{n-m}{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x_j}\right)^2}. \qquad (2.14)$$

Обчислення статистики Фішера – Снедекора зручно виконувати у вигляді спеціальної таблиці дисперсійного аналізу, макет якої з усіма потрібними формулами представлено у табл. 2.10.

Таблиця 2.10

Макет таблиці дисперсійного аналізу

Джерело варіації	Сума квадратів	Ступінь вільності	Середній квадрат	F-критерій (тест F)
Між групами	$\sum_{j=1}^{m}\left(\overline{x_i}-\overline{x}\right)^2\cdot n_j$	m—1	$\dfrac{\sum_{j=1}^{m}\left(\overline{x_i}-\overline{x}\right)^2\cdot n_j}{m-1}$	$F=\dfrac{\dfrac{\sum_{j=1}^{m}\left(\overline{x_i}-\overline{x}\right)^2\cdot n_j}{m-1}}{\dfrac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x_j}\right)^2}{n-m}}$
Всередині груп (випадкова помилка)	$\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x_j}\right)^2$	n—m	$\dfrac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x_j}\right)^2}{n-m}$	

F-критерій (2.14), за допомогою якого ухвалюють рішення про верифікацію (2.10), ґрунтується на розподілі Фішера–Снедекора з кількістю ступенів вільності, що стоять у знаменнику ($\nu_1=m-1$) та чисельнику ($\nu_2=n-m$) формули (2.14).

Найчастіше рівень істотності під час перевірки подібних гіпотез становить $\alpha=0{,}05$, або ж $\alpha=0{,}001$, коли зростає *ціна помилки*, а отже, й відповідальність дослідника. Для таких випадків критичні значення F-статистики за кількості ступенів вільності ν_1 і ν_2 наведено в табл. 2.9. Для випадків з іншими значеннями ступенів вільності чи рівнів істотності табличне значення F-критерію доцільно визначати за допомогою функції Excel FРАСПОБР, синтаксис якої подано для випадку дисперсійного аналізу двох

сукупностей. Критичний простір у такому разі завжди є правобічним. Якщо F-статистика, обчислена за вибірками, перевищить критичне значення $F^*_{\alpha;\nu_1=m-1;\ \nu_2=n-m}$, нуль-гіпотезу про рівність середнього у всіх k сукупностей відкидають й обирають альтернативну:

$$F > F_{\alpha;\nu_1;\nu_2}{}^* \Rightarrow H_0 \Rightarrow H_1(i).$$

Приклад 2.10. Розглянемо приклад верифікації гіпотези про рівність середніх декількох вибірок, проаналізувавши тривалість експонування оголошення про винаймання одно-, дво- та трикімнатних квартир.

Перевірятимемо на рівні істотності $\alpha = 0{,}05$ нуль-гіпотезу про незалежність періоду часу з моменту публікації оголошення до моменту укладання договору купівлі-продажу житла від типу квартири, тобто

$$\boldsymbol{H_0:}\ \bar{x}_1 = \bar{x}_2 = \bar{x}_3 .$$

Відповідно альтернативна гіпотеза констатуватиме брак рівності середнього строку експонування житлових приміщень різного типу:

$$\boldsymbol{H_1:}\ \textit{не правильно, що}\ \bar{x}_1 = \bar{x}_2 = \bar{x}_3 .$$

У результаті верифікації нуль-гіпотези буде встановлений вплив такого фактора, як тип квартири, на тривалість пошуку покупців житла. З метою перевірки гіпотез виконаємо дисперсійний аналіз трьох однакових 5-елементних вибірок (m=3, $n_1=n_2=n_3=5$, n=15), сформованих на основі вивчення бази даних одного з інтернет-порталів нерухомості (табл. 2.11).

Таблиця 2.11

Тривалість експонування житла за даними Інтернет-порталу «www.xxx___xxx.com»

Тип квартири	однокімнатні	двокімнатні	трикімнатні
Спостереження № 1	160	150	250
Спостереження № 2	260	200	400
Спостереження № 3	190	270	220
Спостереження № 4	240	300	350
Спостереження № 5	200	200	300
	210	**224**	**304**
Середній строк експонування	$\bar{x}_1 = \frac{160+260+190+240+200}{5}$	$\bar{x}_2 = \frac{150+200+270+300+200}{5}$	$\bar{x}_3 = \frac{250+400+220+350+300}{5}$

У цій таблиці, разом з даними щодо тривалості експозиції 15 приміщень (по п’ять одно-, дво- та трикімнатних квартир), наведено й середній строк

експонування кожного типу квартир, тобто **групові середні** з відповідними позначеннями ($\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$). **Загальна середня може буде обчислена як середня з групових середніх,** і, відповідно до показників табл. 2.11, вона дорівнюватиме:

$$\bar{x} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{ij}}{\sum_{j=1}^{m} n_j} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{ij}}{n} = \frac{\sum_{j=1}^{m} \bar{x}_j}{m} = \frac{210 + 224 + 304}{3} = 246 \text{ днів}.$$

На підставі обчислених середніх можна визначити дисперсійні суми квадратів:

– суму, що відображає варіацію між групами:

$$\sum_{j=1}^{m}\left(\bar{x}_i - \bar{x}\right)^2 \cdot n_j = (210 - 246)^2 \cdot 5 + (224 - 246)^2 \cdot 5 + (304 - 246)^2 \cdot 5 = 25720.$$

При цьому кількість ступенів вільності дорівнює 2, що на 1 менше за кількість груп (3—1);

– суму, яка визначає варіації всередині груп, тобто залишкову варіацію $\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2\right)$, зручніше обчислити за допомогою допоміжної, робочої таблиці (табл. 2.12). Підсумовуючи дисперсійні суми у кожній з груп, одержимо 42240(=6400+14520+21320).

Кількість ступенів вільності для випадкової помилки становить 12, що є різницею між загальним обсягом всіх вибірок та кількістю вибірок (15–3).

Таблиця 2.12

Робоча таблиця для обчислення залишкової дисперсії

Тип квартири	Однокімнатні		Двокімнатні		Трикімнатні	
	строк експозиції	квадрат відхилення від групової середньої	строк експозиції	квадрат відхилення від групової середньої	строк експозиції	квадрат відхилення від групової середньої
Спостереження № 1	160	$(160-210)^2$=**2500**	150	$(150-224)^2$=**5476**	250	$(250-304)^2$=**2916**
Спостереження № 2	260	$(260-210)^2$=**2500**	200	$(200-224)^2$=**576**	400	$(400-304)^2$=**9216**
Спостереження № 3	190	$(190-210)^2$=**400**	270	$(270-224)^2$=**2116**	220	$(220-304)^2$=**7056**
Спостереження № 4	240	$(240-210)^2$=**900**	300	$(300-224)^2$=**5776**	350	$(350-304)^2$=**2116**
Спостереження № 5	200	$(200-210)^2$=**100**	200	$(200-224)^2$=**576**	300	$(300-304)^2$=**16**
Дисперсійна сума	**6400**		**14520**		**21320**	

Подальше обчислення статистики Фішера–Снедекора представлено у таблиці дисперсійного аналізу (табл. 2.13).

Таблиця 2.13

Результати дисперсійного аналізу тривалості експонування квартир

Джерело варіації	Сума квадратів	Ступінь вільності	Середній квадрат	F-критерій (тест F)
Між групами	$\sum_{j=1}^{m}\left(\overline{x_i}-\overline{x}\right)^2 \cdot n_j = 25720$	m-1=2	$\frac{\sum_{j=1}^{m}\left(\overline{x_i}-\overline{x}\right)^2 \cdot n_j}{m-1} = \frac{25720}{2} = 12860$	$F = \frac{\frac{\sum_{j=1}^{m}\left(\overline{x_i}-\overline{x}\right)^2 \cdot n_j}{m-1}}{\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x_j}\right)^2}{n-m}}$ $F = \frac{12860}{3520} = 3{,}653$
Всередині груп (випадкова помилка)	$\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x_j}\right)^2 = 42240$	n-m=12	$\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x_j}\right)^2}{n-m} = \frac{42240}{12} = 3520$	

Критичне значення F-критерію згідно з табл. 2.7 становить $F^*_{\alpha=0.05;\nu_1=2;\nu_2=12} = 3{,}885$ (такий самий результат можна отримати, ввівши до клітинки Excel формулу =F.INV.RT(0.05;2;12) (=FРАСПОБР(0.05;2;12) для ранньої версії програми)). Оскільки F-статистика, обчислена за вибірками, є меншою за критичне значення $F^*_{(0{,}05;2;12)}$, нуль-гіпотеза про рівність строку експонування квартир незалежно від кількості кімнат підтверджується:

$$3{,}653 < 3{,}885 \Rightarrow F < F_{0{,}05;2;12}{}^* \Rightarrow \mathrm{H}_0(\mathrm{i}).$$

Таким чином, на рівні значущості $\alpha = 0{,}05$ можна стверджувати, що строк експонування квартир на інтернет-порталі www.xxx___xxx.com не залежить від кількості кімнат у квартирі. Стверджуючи це, можна припускатись помилки лише в п’ятьох випадках із 100.

ЗАВДАННЯ ДЛЯ САМОСТІЙНОГО ОПРАЦЮВАННЯ МАТЕРІАЛУ

Увага! У визначенні числових показників для розрахунково-аналітичних завдань слід мати на увазі:

- **h*** – **остання цифра** номера вашого паспорта;
- **g*** – **передостання цифра** номера вашого паспорта.

Якщо біля числового показника немає буквених доданків чи співмножників, **такий показник є спільним для всіх варіантів.**

1. З метою вивчення попиту на розчинник для фарб були проаналізовані обсяги продажу 30+g* торгових точок протягом 15 робочих днів. Середньоденний обсяг продажу в одній торговій точці становив 30–h* пляшок розчиннику для фарб зі стандартним відхиленням 2+g* пляшок. Обчисліть обсяг вибірки, тобто загальну кількість спостережень. Чи є вибірка великою, чи малою? На рівні значущості (1+h*)% перевірте гіпотезу про те,

що середньоденний попит в одній торговій **точці** становить 25 пляшок розчиннику для фарб.

2. Гарантійний строк служби батарейки – щонайменше 1 000 годин. З вибірки 12+g* батарейок встановлено, що середній строк служби становить 1002 год зі стандартним відхиленням 30+g* год. Чи правдивою є обіцянка виробника? Якою є критична область – одно- чи двобічною? Обґрунтуйте своє рішення. Для верифікації гіпотези оберіть рівень надійності згідно з вашим варіантом за табл. 2.14.

Таблиця 2.14

Рівень значущості для верифікації гіпотез про середнє значення

Варіант, h*	**0;4**	**1;5**	**2;8**	**3;9**	**6;7**
Імовірність помилки (однобічна область, $\frac{\alpha}{2}$)	**0,001**	**0,002**	**0,01**	**0,02**	**0,05**

3. За наведеними в табл. 2.15 даними виконайте такі розрахунки.

- Проаналізуйте обсяг виробітку двох бригад будівельників-мулярів за одну зміну. На основі коефіцієнта варіації визначте, у якій бригаді більш однорідний склад робітників за рівнем продуктивності праці.
- Чи можна стверджувати, що змінний виробіток у першій і другій бригадах є однаковим?

Таблиця 2.15

Вихідні дані для перевірки статистичних гіпотез

Бригада №1			**Бригада №2**			**Бригада №3 (студенти)**		
Робітник	**Стаж роботи, років**	**Змінний виробіток, $м^3$ цегляної кладки**	**Робітник**	**Стаж роботи, років**	**Змінний виробіток, $м^3$ цегляної кладки**	**Робітник**	**Стаж робо-ти, років**	**Змінний виробіток, $м^3$ цегляної кладки**
Воробйов	5	22	Азазело	5	22	Береза	3	22
Дроздов	5,5	16+1,5•h*	Бегемот	4	16+h*	Верба	2	16–0,5•h*
Снєгірьов	6	26+h*	Воланд	3	26—0,5•h*	Дубовий	4,5	26–1,5•h*
Соловйов	4,5	20+0,5•g*	Коровьєв	4	20+1,5•g*	Осика	2,5	20–g*
Чижов	7	28	Майстер	7	28	Тополя	4	28

Надійність рішення з верифікації обрати за вашим варіантом завдання і табл. 2.16.

Таблиця 2.16

Рівень значущості для верифікації гіпотез про середнє значення у двох вибірках

Варіант, h*	**1;6**	**2;7**	**3;8**	**4;9**	**5;0**
імовірність помилки (однобічна область, α)	**0,1**	**0,05**	**0,025**	**0,005**	**0,0005**

- Чи можна стверджувати, що змінний виробіток в середньому однаковий у всіх трьох бригадах? Надійність рішення з верифікації оберіть за вашим варіантом завдання й табл. 2.17.

Таблиця 2.17

Рівень значущості для дисперсійного аналізу

Варіант, h*	1;4;7	2;5;8;0	3;6;9
Імовірність помилки, α	0,1	0,05	0,025

- Чи можна стверджувати, що немає розбіжностей у показниках дисперсії виробітку у бригадах № 2 та № 3? Надійність рішення з верифікації оберіть за вашим варіантом завдання й табл. 2.18

Таблиця 2.18

Рівень значущості для верифікації гіпотез про рівність дисперсій двох вибірок

Варіант, h*	2;5;8;0	3;6;9	1;4;7
Імовірність помилки, α	0,1	0,05	0,025

4. Чи справедливим є твердження, що менш ніж 35% покупців дізнаються про акційні ціни на товари за допомогою Інтернету, якщо з 250−5•g* осіб Інтернет був джерелом інформації для 75—h* осіб? Якою є критична область − одно- чи двобічною? Обґрунтуйте своє рішення. Для перевірки гіпотези взяти рівень значущості (0,5+h*)%.

5. Менеджери мережі магазинів з продажу освітлювальної техніки вирішили з'ясувати, як впливає музичний супровід на обсяг продажу світлодіодних ламп. В одному магазині, де систематично протягом місяця вмикали джаз, з 2000+20•g* осіб, що цікавились світлодіодними лампами, покупки зробили 500+10•h*. В іншому магазині під акомпанемент попси з 3000−20•g* осіб, що цікавились світлодіодними лампами, пішли з магазину з покупками 600. З'ясувати, чи впливає музичний супровід на рішення придбати світлодіодні лампи (рівень значущості − 0,05 та 0,01).

6. Чи розподіл видатків на ремонт житла домогосподарств є відповідним нормальному? Які параметри нормального розподілу слід перевірити відповідно до вашого варіанта завдання? Самостійно згрупуйте вхідні дані за інтервалами відповідно до формули Стерджеса.

g* − непарне (1,3,5,7,9)										g* − парне (0,2,4,6,8)									
h*=1	h*=2	h*=3	h*=4	h*=5	h*=6	h*=7	h*=8	h*=9	h*=0	h*=1	h*=2	h*=3	h*=4	h*=5	h*=6	h*=7	h*=8	h*=9	h*=0
1479	1508	1537	1566	1595	1624	1653	1682	1711	1450	2907	2964	3021	3078	3135	3192	3249	3306	3363	2850
1734	1768	1802	1836	1870	1904	1938	1972	2006	1700	1326	1352	1378	1404	1430	1456	1482	1508	1534	1300
1530	1560	1590	1620	1650	1680	1710	1740	1770	1500	2346	2392	2438	2484	2530	2576	2622	2668	2714	2300
1275	1300	1325	1350	1375	1400	1425	1450	1475	1250	1887	1924	1961	1998	2035	2072	2109	2146	2183	1850
2550	2600	2650	2700	2750	2800	2850	2900	2950	2500	1326	1352	1378	1404	1430	1456	1482	1508	1534	1300
1989	2028	2067	2106	2145	2184	2223	2262	2301	1950	1938	1976	2014	2052	2090	2128	2166	2204	2242	1900
1938	1976	2014	2052	2090	2128	2166	2204	2242	1900	1632	1664	1696	1728	1760	1792	1824	1856	1888	1600
1377	1404	1431	1458	1485	1512	1539	1566	1593	1350	1122	1144	1166	1188	1210	1232	1254	1276	1298	1100
3009	3068	3127	3186	3245	3304	3363	3422	3481	2950	1183	1206	1230	1253	1276	1299	1322	1346	1369	1160
867	884	901	918	935	952	969	986	1003	850	1887	1924	1961	1998	2035	2072	2109	2146	2183	1850
1428	1456	1484	1512	1540	1568	1596	1624	1652	1400	561	572	583	594	605	616	627	638	649	550
2244	2288	2332	2376	2420	2464	2508	2552	2596	2200	1581	1612	1643	1674	1705	1736	1767	1798	1829	1550
2040	2080	2120	2160	2200	2240	2280	2320	2360	2000	1836	1872	1908	1944	1980	2016	2052	2088	2124	1800
1836	1872	1908	1944	1980	2016	2052	2088	2124	1800	1785	1820	1855	1890	1925	1960	1995	2030	2065	1750
2397	2444	2491	2538	2585	2632	2679	2726	2773	2350	1275	1300	1325	1350	1375	1400	1425	1450	1475	1250
1122	1144	1166	1188	1210	1232	1254	1276	1298	1100	1581	1612	1643	1674	1705	1736	1767	1798	1829	1550
1989	2028	2067	2106	2145	2184	2223	2262	2301	1950	1683	1716	1749	1782	1815	1848	1881	1914	1947	1650
1734	1768	1802	1836	1870	1904	1938	1972	2006	1700	2601	2652	2703	2754	2805	2856	2907	2958	3009	2550
2601	2652	2703	2754	2805	2856	2907	2958	3009	2550	1836	1872	1908	1944	1980	2016	2052	2088	2124	1800

1887	1924	1961	1998	2035	2072	2109	2146	2183	1850	1326	1352	1378	1404	1430	1456	1482	1508	1534	1300
2448	2496	2544	2592	2640	2688	2736	2784	2832	2400	1224	1248	1272	1296	1320	1344	1368	1392	1416	1200
2652	2704	2756	2808	2860	2912	2964	3016	3068	2600	2499	2548	2597	2646	2695	2744	2793	2842	2891	2450
g* – непарне (1,3,5,7,9)										**g* – парне (0,2,4,6,8)**									
h*=1	**h*=2**	**h*=3**	**h*=4**	**h*=5**	**h*=6**	**h*=7**	**h*=8**	**h*=9**	**h*=0**	**h*=1**	**h*=2**	**h*=3**	**h*=4**	**h*=5**	**h*=6**	**h*=7**	**h*=8**	**h*=9**	**h*=0**
1989	2028	2067	2106	2145	2184	2223	2262	2301	1950	2193	2236	2279	2322	2365	2408	2451	2494	2537	2150
2193	2236	2279	2322	2365	2408	2451	2494	2537	2150	2193	2236	2279	2322	2365	2408	2451	2494	2537	2150
2295	2340	2385	2430	2475	2520	2565	2610	2655	2250	2244	2288	2332	2376	2420	2464	2508	2552	2596	2200
1428	1456	1484	1512	1540	1568	1596	1624	1652	1400	2499	2548	2597	2646	2695	2744	2793	2842	2891	2450
2040	2080	2120	2160	2200	2240	2280	2320	2360	2000	2142	2184	2226	2268	2310	2352	2394	2436	2478	2100
2346	2392	2438	2484	2530	2576	2622	2668	2714	2300	2193	2236	2279	2322	2365	2408	2451	2494	2537	2150
1683	1716	1749	1782	1815	1848	1881	1914	1947	1650	1734	1768	1802	1836	1870	1904	1938	1972	2006	1700
1326	1352	1378	1404	1430	1456	1482	1508	1534	1300	1122	1144	1166	1188	1210	1232	1254	1276	1298	1100
1785	1820	1855	1890	1925	1960	1995	2030	2065	1750	1887	1924	1961	1998	2035	2072	2109	2146	2183	1850
2652	2704	2756	2808	2860	2912	2964	3016	3068	2600	1836	1872	1908	1944	1980	2016	2052	2088	2124	1800
867	884	901	918	935	952	969	986	1003	850	1479	1508	1537	1566	1595	1624	1653	1682	1711	1450
2295	2340	2385	2430	2475	2520	2565	2610	2655	2250	1632	1664	1696	1728	1760	1792	1824	1856	1888	1600
2703	2756	2809	2862	2915	2968	3021	3074	3127	2650	2346	2392	2438	2484	2530	2576	2622	2668	2714	2300
1887	1924	1961	1998	2035	2072	2109	2146	2183	1850	1275	1300	1325	1350	1375	1400	1425	1450	1475	1250
2142	2184	2226	2268	2310	2352	2394	2436	2478	2100	2652	2704	2756	2808	2860	2912	2964	3016	3068	2600
2193	2236	2279	2322	2365	2408	2451	2494	2537	2150	2193	2236	2279	2322	2365	2408	2451	2494	2537	2150
1530	1560	1590	1620	1650	1680	1710	1740	1770	1500	1020	1040	1060	1080	1100	1120	1140	1160	1180	1000
1734	1768	1802	1836	1870	1904	1938	1972	2006	1700	2040	2080	2120	2160	2200	2240	2280	2320	2360	2000

Розділ 3

СТАТИСТИЧНІ МЕТОДИ ВИВЧЕННЯ ЗВ'ЯЗКУ МІЖ ЯВИЩАМИ ТА ПРОЦЕСАМИ БУДІВНИЦТВА І УПРАВЛІННЯ ОБ'ЄКТАМИ НЕРУХОМОСТІ

3.1. КОРЕЛЯЦІЙНИЙ ЗВ'ЯЗОК

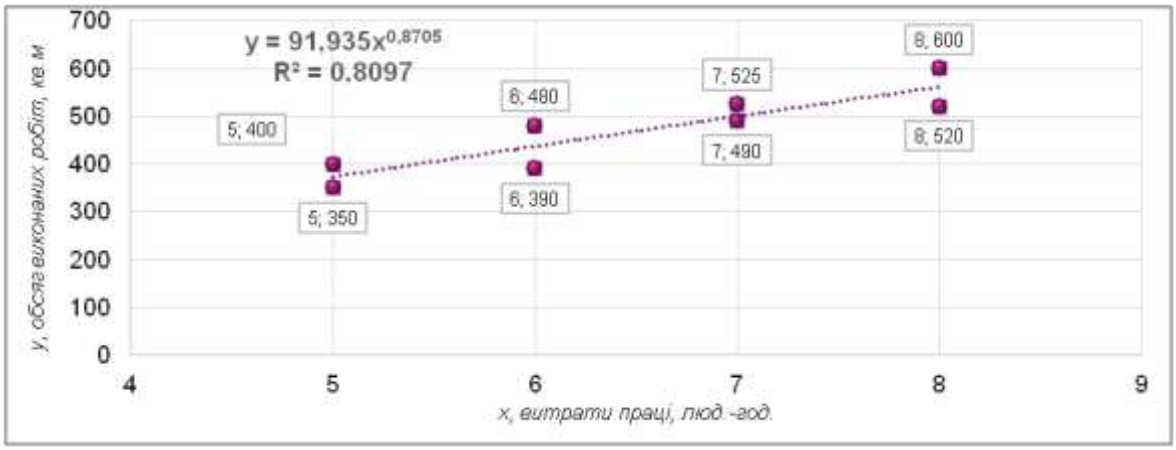
3.1.1. Основні теоретичні відомості

Взаємозв'язок між двома ознаками, x та y характеризується такими поняттями, як **форма** та **щільність**. Їх зручно проілюструвати графічно (рис.3.1, *а*, *б*).

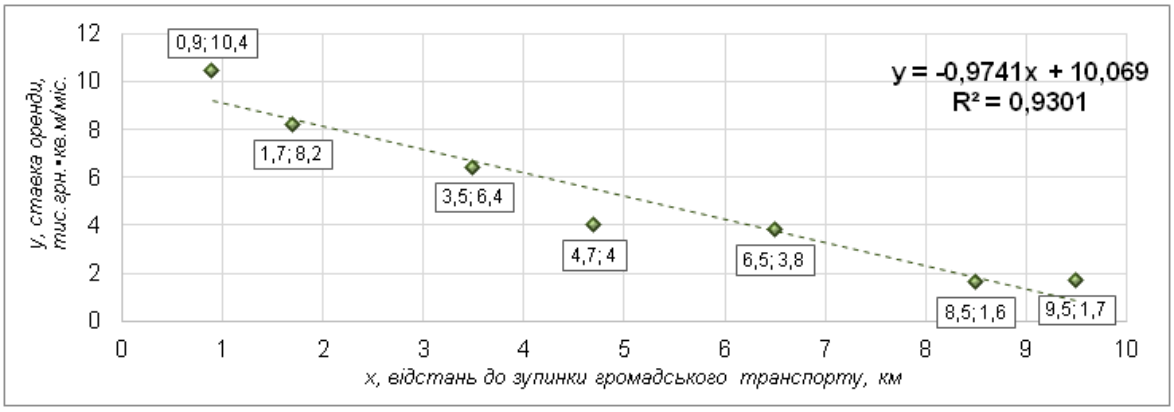
Форма відображає конфігурацію хмари точок на графіку, а щільність – міру розсіювання фактичних спостережень-точок навколо теоретичної лінії. Ця лінія (рис. 3.1) визначається певною функцією, що апроксимує емпіричні дані теоретичним рівнянням.

На рис. 3.1, *а* представлено пряму залежність між ознаками *x* та *y*, оскільки хмара точок прямує до верхнього правого кута діаграми, тоді як рис. 3.1, *б* ілюструє обернену залежність, адже точки на графіку «тяжіють» до правого нижнього кута. Крім того, штрихові лінії на графіках визначаються

різними рівняннями: ступеневим (рис.3.1, *а*) й лінійним (рис. 3.1, *б*). Такі теоретичні лінії мають також назву **ліній регресії**.



а – зростання нелінійної залежності обсягу виконаних робіт (у) від витрат праці (х)



б – спадна лінійна залежність ставки оренди комерційних приміщень (у) від відстані до найближчої зупинки громадського транспорту (х)

Рис. 3.1. Графічне подання взаємозв'язку між двома ознаками

На рис. 3.1, *б* більшість точок майже лежать на пунктирній теоретичній лінії, тимчасом як на графіку 3.1, *а* більшість точок знаходяться на віддалі від теоретичної лінії. Це пояснюється різною щільністю зв'язку між ознаками: що ближче точки прилягають до графіка, то щільність зв'язку вища. Одним із способів визначення щільності зв'язку є розрахунок **коефіцієнта достовірності апроксимації** теоретичним рівнянням фактичних даних (R-квадрат, R^2). Значення цього коефіцієнта представлено на графіках, причому на рис. 3.1, *б* R^2 більший: 0,9301>0,8101. Отже, залежність орендної ставки від розміщення об'єктів є більш щільною порівняно із залежністю обсягу виконаних робіт від затрат праці.

Методику розрахунку регресійних рівнянь та перевірки достовірності їхньої апроксимації фактичних зв'язків між ознаками викладено далі, оскільки таким обчисленням передує перевірка ознак на узгодженість їхніх змін. Адже

часто буває недоцільно будувати регресійні рівняння через брак або слабкість зв'язку між ознаками, тобто низьку кореляцію.

Кореляційно-регресійний аналіз припускає, що ознаки **x** та **y** підпорядковано сумісному нормальному розподілу.

Корельованість, або узгодженість, зміни ознак виявляється в тому, що більшим за модулем відхиленням від середнього значення $\bar{x}$ відповідною є й вища варіація ознаки $\bar{y}$. Для виміру такої узгодженості використовують змішаний момент 2-го порядку, тобто коваріацію ознак **x** та **y**. Позначають коваріацію так: **cov(x,y)**.

Коефіцієнт коваріації обчислюють за формулою:

$$\operatorname{cov}(x,y)=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\bullet\left(y_i-\bar{y}\right)}{n}. \tag{3.1}$$

Коефіцієнт коваріації може бути визначений і як «виправлений», коли в знаменнику замість обсягу вибірки *n* буде кількість ступенів вільності, *n*–1. Втім, цей показник частіше використовують як проміжний, далі показано, що вибір кількості ступенів вільності як знаменника не є принциповим.

Оскільки чисельник формули (3.1) містить не квадрати, а добутки різниць, коваріація може бути як додатним, так і від'ємним числом:

- **додатне** значення **коваріації** свідчить про прямий зв'язок, тобто у міру зростання значення ознаки x збільшується й значення ознаки y або ж значення обох ознак зменшуються;
- **від'ємне** значення **коваріації** буває тоді, коли зв'язок між x та y є оберненим. При цьому у разі зростання значення ознаки x значення ознаки y зменшується, і навпаки.

Розкриваючи дужки (3.1) та вдаючись до елементарних алгебраїчних перетворень, можна отримати формулу «сирого розрахунку» коефіцієнта коваріації. Вона легка для запам'ятовування та прискорює обчислення у разі великих масивів даних про ознаки x та y:

$$\operatorname{cov}(x,y)=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\bullet\left(y_i-\bar{y}\right)}{n}=\frac{\sum_{i=1}^{n}\left(x_i\bullet y_i\right)-\bar{x}\bullet\sum_{i=1}^{n}y_i-\bar{y}\bullet\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\left(\bar{x}\bullet\bar{y}\right)}{n}=$$

$$=\underbrace{\frac{\sum_{i=1}^{n}x_i\bullet y_i}{n}}_{\overline{x\bullet y}}-\underbrace{\left(\bar{x}\bullet\frac{\sum_{i=1}^{n}y_i}{n}+\bar{x}\bullet\frac{\sum_{i=1}^{n}y_i}{n}\right)}_{2\bullet\bar{x}\bullet\bar{y}}+\underbrace{\frac{\sum_{i=1}^{n}\left(\bar{x}\bullet\bar{y}\right)}{n}}_{\bar{x}\bullet\bar{y}}=$$

$$=\overline{x\bullet y}-2\bullet\bar{x}\bullet\bar{y}+\bar{x}\bullet\bar{y}=\overline{x\bullet y}-\bar{x}\bullet\bar{y}$$

$$\Downarrow$$

$$\operatorname{cov}(x,y)=\overline{x\bullet y}-\bar{x}\bullet\bar{y}$$

$$\mathrm{cov}(x,y) = \overline{x \bullet y} - \bar{x} \bullet \bar{y}, \qquad (3.2)$$

де $\overline{x \bullet y}$ – середнє значення добутку ознак x_i та y_i для кожного спостереження.

Приклад 3.1. За даними рис. 3.1 визначте коваріацію між обсягами виконаних робіт та затратами праці й ставкою оренди та віддаленістю об'єкта від зупинки громадського транспорту. Для другої пари показників вибірка вихідних даних є меншою, тому обчислимо коваріацію безпосередньо за допомогою формули 3.1. Перед її застосуванням потрібно розрахувати середні значення кожної з ознак:

$$\bar{x} = \frac{0{,}9 + 1{,}7 + 3{,}5 + 4{,}7 + 6{,}5 + 8{,}5 + 9{,}5}{7} = \frac{35{,}3}{7} = 5{,}04;$$

$$\bar{y} = \frac{10{,}4 + 8{,}2 + 6{,}4 + 4 + 3{,}8 + 1{,}7 + 1{,}6}{7} = \frac{36{,}1}{7} = 5{,}16;$$

$$\mathrm{cov}(x,y) = \frac{(0{,}9 - 5{,}04) \bullet (10{,}4 - 5{,}16) + (1{,}7 - 5{,}04) \bullet (8{,}2 - 5{,}16) + (3{,}5 - 5{,}04) \bullet (6{,}4 - 5{,}16)}{7} +$$

$$+ \frac{(4{,}7 - 5{,}04) \bullet (4 - 5{,}16) + (6{,}5 - 5{,}04) \bullet (3{,}8 - 5{,}16) + (8{,}5 - 5{,}04) \bullet (1{,}7 - 5{,}16) + (9{,}5 - 5{,}04) \bullet (1{,}6 - 5{,}16)}{7};$$

$$\mathrm{cov}(x,y) = \frac{-1{,}92 - 10{,}17 - 15{,}41 - 1{,}98 - 21{,}72 - 12{,}3 + 0{,}4}{7} = \frac{-63{,}1}{7} = -9{,}01.$$

Отриманий коефіцієнт коваріації є безрозмірною величиною, його від'ємне значення свідчить про зворотний зв'язок між x та y: відстань від зупинки та ставка оренди змінюються різноспрямовано (рис.3.1, *б*), Дорожчою оренда є на тих об'єктах торгівлі, що розміщені поблизу людних місць – зупинок транспорту. Натомість об'єкти, діставатись до яких довго, дають менший виторг, вони непривабливі і для орендарів, а тому їхнє винаймання коштує дешевше.

Сім спостережень – це мала вибірка, та все ж розрахунок виявився громіздким, дріб з усіма добутками різниць не вмістився в межі сторінки. Для більш компактного запису можна скористатися формулою «сирого розрахунку» (3.2.). Її використання, крім середніх значень ознак, потребує усереднення добутків кожної пари ознак.

$$\overline{x \bullet y} = \frac{0{,}9 \bullet 10{,}4 + 1{,}7 \bullet 8{,}2 + 3{,}5 \bullet 6{,}4 + 4{,}7 \bullet 4 + 6{,}5 \bullet 3{,}8 + 8{,}5 \bullet 1{,}7 + 9{,}5 \bullet 1{,}6}{7};$$

$$\overline{x \bullet y} = \frac{9{,}36 + 13{,}94 + 22{,}4 + 18{,}8 + 24{,}7 + 24{,}7 + 16{,}15 + 13{,}6}{7} = \frac{118{,}95}{7} \approx 17$$

$$\Downarrow$$

$$\overline{x \bullet y} - \bar{x} \bullet \bar{y} = 17 - 5{,}04 \bullet 5{,}16 = -9{,}01.$$

Результати використання формул (3.1) та (3.2), як і слід було сподіватися, збіглися.

Звісно, такі розрахунки можна виконувати не рядком, а у вигляді таблиці.

Для визначення коваріації ознак з рис. 3.1.*а* про обсяги робіт та затрати праці розрахунки виконано у табл. 3.1.

Результати обчислення коваріації затрат праці та обсягів виконаних робіт з улаштування покрівлі, за інформацією будівельних підприємств ТОВ «Аркадос» і ТОВ «Коймаран» (див. приклад 1.4, рис. 3.1, *а*), такі:

– традиційна формула (3.1):

$$cov(x,y)=\frac{\sum_{i=1}^{8}\left(x_i-\bar{x}\right)\bullet\left(y_i-\bar{y}\right)}{n}=\frac{627{,}5}{8}=78{,}4375>0;$$

– формула «сирого розрахунку» (3.2):

$$\mathrm{cov}(x,y)=\overline{x\bullet y}-\bar{x}\bullet\bar{y}=3129{,}38-6{,}5\bullet 469{,}375=78{,}4375>0.$$

Результати обох формул збіглися, показник коваріації додатний, отже, зв'язок між витратами праці та обсягом виконаних робіт прямий: що більше людино-годин затрачено на роботу, тим більшим є обсяг виконаних робіт.

Таблиця 3.1

Вихідні дані та проміжні розрахунки для обчислення коефіцієнта коваріації

Порядковий номер спостереження	**Витрати праці, x_i**	**Обсяг виконаних робіт із улаштування покрівлі, y_i**	**Добуток значень спостережень**	**Відхилення від середнього факторної ознаки** $x_i-\bar{x}=$ $=x_i-6{,}5$	**Відхилення від середнього результаційної ознаки** $y_i-\bar{y}=$ $=y_i-469{,}375$	**Добуток відхилень ознак від середніх** $\left(x_i-\bar{x}\right)\bullet\left(y_i-\bar{y}\right)=$ $=\left(x_i-6{,}5\right)\bullet\left(469{,}375\right)$
1	5	350	1750	-1.5	-119.38	179.063
2	6	480	2880	-0.5	10.625	-5.3125
3	7	525	3675	0.5	55.625	27.8125
4	8	520	4160	1.5	50.625	75.9375
5	8	600	4800	1.5	130.625	195.938
6	6	390	2340	-0.5	-79.375	39.6875
7	5	400	2000	-1.5	-69.375	104.063
8	7	490	3430	0.5	20.625	10.3125
Разом	52	3755	25035	0	0	$\sum_{i=1}^{8}\left(x_i-\bar{x}\right)\bullet\left(y_i-\bar{y}\right)=$ $=627.5$

Середнє	$\bar{x} = 6.5$	$\bar{y} = 469.375$	$\overline{x \bullet y} =$ =3129.38	0	0	cov(x,y)=78.4375

Показник коваріації покладено в основу обчислень решти показників кореляційно-регресійного аналізу, зокрема коефіцієнтів регресійних рівнянь, достовірності апроксимації та коефіцієнта парної кореляції, який усуває основний недолік коефіцієнта коваріації – незіставність результатів обчислень через масштаб значень ознак. Зокрема, для прикладу з рис. 3.1, *а* коваріація – 78,4, а для рис. 3.1, *б* – лише 9,01. Попри різний напрям зв'язку, висновку про щільність зв'язку ознак зробити неможливо.

3.1.2. КОЕФІЦІЄНТ ПАРНОЇ КОРЕЛЯЦІЇ

Для можливості порівнювати щільність зв'язку між різними парами ознак обчислюють коефіцієнт парної кореляції, що являє собою стандартизований коефіцієнт коваріації. Стандартизація полягає в тому, що коефіцієнт коваріації ділять на добуток стандартних відхилень ознак x та y (σ_x, σ_y). Коефіцієнт парної кореляції позначають як cor(x,y) або R(x,y) й обчислюють так:

$$cor(x,y) = R(x,y) = \frac{\mathrm{cov}(x,y)}{\sigma_x \bullet \sigma_y}. \qquad (3.3)$$

Цей коефіцієнт можна обчислювати, використовуючи також відомі формули «сирого розрахунку»:

$$cor(x,y) = \frac{\overline{x \bullet y} - \bar{x} \bullet \bar{y}}{\sqrt{\left(\overline{x^2} - \bar{x}^2\right) \bullet \left(\overline{y^2} - \bar{y}^2\right)}}. \qquad (3.4)$$

Оскільки і чисельник, і знаменник формули (3.3) являють собою усереднені показники, знаменники яких скорочують, то **не має значення, чи будуть використані виправлені** коваріація й дисперсія, чи не будуть виправлені: в обох випадках *n*–1 та n будуть скорочені. Це дає змогу впевнено застосовувати усі формули «сирого розрахунку» без виправлень.

Основні властивості коефіцієнта парної кореляції:

- **коефіцієнт кореляції** – **безрозмірна** величина;
- **коефіцієнт кореляції** – **обмежена** величина, він перебуває в межах від –1 до 1:

$$-1 \le cor(x,y) \le 1 \qquad (3.5)$$

- якщо наявна сувора рівність **модуля коефіцієнта кореляції 1:** $|cor(x,y)| = 1$ **(**тобто **cor(x,y)=1 або cor(x,y)=—1),** зв'язок між *x* та *y* – **лінійний;**

- **що ближче значення модуля коефіцієнта кореляції, то вищою є щільність лінійного зв'язку;**
- якщо **cor(x,y)>0**, зв'язок **прямий**, якщо **cor(x,y)<0**, зв'язок **зворотний**, якщо **cor(x,y)=0**, лінійного зв'язку немає**, проте можливим є нелінійний зв'язок**.

Приклад 3.2. За даними прикладу 3.1. та рис. 3.1 визначте кореляцію між x та y для продуктивності праці та оренди торговельних приміщень.

Для ситуації зі ставкою оренди розрахунок виконаємо без складання таблиць. Використаємо формулу (3.4), для цього знайдемо усереднені квадрати ознак x та y:

$$\overline{x^2} = \frac{0{,}9^2 + 1{,}7^2 + 3{,}5^2 + 4{,}7^2 + 6{,}5^2 + 8{,}5^2 + 9{,}5^2}{7} = \frac{0{,}81 + 2{,}89 + 12{,}25 + 22{,}09 + 42{,}25 + 72{,}25 + 90,}{7} =$$

$$= \frac{242{,}8}{7} = 34{,}69;$$

$$\overline{y^2} = \frac{10{,}4^2 + 8{,}2^2 + 6{,}4^2 + 4^2 + 3{,}8^2 + 1{,}7^2 + 1{,}6^2}{7} = \frac{108{,}16 + 67{,}24 + 40{,}96 + 16 + 11{,}4 + 2{,}89 + 2{,}56}{7} =$$

$$= \frac{349{,}21}{7} = 35{,}6;$$

$$R(x,y) = cor(x,y) = \frac{\mathrm{cov}(x,y)}{\sigma_x \bullet \sigma_y} = \frac{-9{,}01}{\sqrt{(34{,}69 - 5{,}04^2) \bullet (35{,}6 - 5{,}16^2)}} =$$

$$= \frac{-9{,}01}{\sqrt{9{,}28 \bullet 8{,}97}} = \frac{-9{,}01}{\sqrt{83{,}36}} = \frac{-9{,}01}{9{,}13} = -0{,}987.$$

Отриманий результат від'ємний та за модулем близький до 1: $|-0{,}987| \approx 1$. Отже, спостерігається майже прямолінійна спадна залежність між вартістю оренди та віддаленістю торговельної нерухомості від зупинок громадського транспорту.

Для прикладу про затрати праці та обсяги робіт додаткові дані обчислимо у вигляді таблиці (табл. 3.2)

Таблиця 3.2

Додаткові проміжні розрахунки для обчислення коефіцієнта парної кореляції

Порядковий номер спостереження	**Витрати праці, x_i**	**Обсяг виконаних робіт із улаштування покрівлі, y_i**	**Відхилення від середнього факторної ознаки** $x_i - \overline{x} = x_i - 6{,}5$	**Квадрат відхилення від середнього факторної ознаки** $(x_i - \overline{x})^2 = (x_i - 6{,}5)^2$	**Відхилення від середнього результаційної ознаки** $y_i - \overline{y} = y_i - 469{,}375$	**Квадрат відхилення від середнього результаційної ознаки** $(y_i - \overline{y})^2 = (y_i - 469{,}375)^2$
1	5	350	-1.5	2,25	-119.375	14250.391
2	6	480	-0.5	0,25	10.625	112.891
3	7	525	0.5	0,25	55.625	3094.141

4	8	520	1.5	2,25	50.625	2562.891
5	8	600	1.5	2,25	130.625	17062.891
6	6	390	-0.5	0,25	-79.375	6300.391
7	5	400	-1.5	2,25	-69.375	4812.891
8	7	490	0.5	0,25	20.625	425.391
Разом	52	3755	0	10	0	48621.875
Середнє	$\overline{x} = 6.5$	$\overline{y} = 469.375$	0	$\sigma_x^2 =$ **1,25**	0	$\sigma_y^2 =$ **6077.734**

За формулою 3.3 одержимо;

$$R(x,y) = cor(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \bullet \sigma_y} = = \frac{78{,}4375}{\sqrt{1{,}25 \bullet 6077{,}734}} = \frac{78{,}4375}{\sqrt{7957{,}17}} = \frac{78{,}4375}{87{,}16} = 0{,}90 .$$

Отриманий результат – додатний і також близький до 1 за модулем: $|0{,}90| \approx 1$. Отже, спостерігає щільна пряма, майже лінійна залежність між витратами труда та обсягом виконаних робіт.

3.1.3. Перевірка значущості коефіцієнта парної кореляції

Оскільки щільність зв'язку часто визначають на малих вибірках, постає проблема перевірки статистичної значущості отриманих результатів. Тобто необхідно встановити, чи виявлена щільність зв'язку буде такою самою і в інших вибірках за заданого рівня імовірності помилки α. Для цього перевіряють статистичну гіпотезу:

H_0: **Між ознаками *x* та *y* немає кореляційного зв'язку**, тобто нуль-гіпотезу формалізуємо: **H_0:R(x,y)=0.**

Тоді альтернативну гіпотезу сформулюємо так:

H_1: **Між ознаками x та y є кореляційний зв'язок будь-якого напрямку**, тобто альтернативну гіпотезу формалізуємо: **H_1:R(x,y)** $\neq$ **0.** Гіпотеза є двобічною, бо містить знак «$\neq$».

Гіпотезу перевіряють за розрахунком *t*-критерію за формулою:

$$t = \sqrt{\frac{(R(x,y))^2 \bullet (n-2)}{1-(R(x,y))^2}}, \qquad (3.6)$$

де $(R(x,y))^2$ – квадрат коефіцієнта парної кореляції;

$n-2=\nu$ – кількість ступенів вільності. Від *n* віднімаємо не 1, а 2, оскільки в розрахунках кореляції використовують два середніх для кожної з ознак – *x* та *y*, і на кожну припадає по одному ступеню вільності.

Нуль-гіпотезу відкидають і беруть альтернативну, якщо розрахунковий t-критерій перевищить t-критерій табличний $t^*_{\frac{\alpha}{2},\nu=n-2}$. Табличний t*-критерій визначають за кількості ступенів вільності $\nu = n-2$ (як й у формулі (3.6)) та довірчою імовірністю $\frac{\alpha}{2}$ (бо критична область є двобічною):

$$|t| > \left| t^*_{\frac{\alpha}{2};\nu=n-2} \right| \Rightarrow \text{H}_0 \Rightarrow \text{H}_1(\text{i}). \qquad (3.7)$$

Приклад 3.3. За даними прикладу 3.2 перевіримо значущість отриманих коефіцієнтів кореляції. При цьому для продуктивності праці (рис. 3.1, *а*) імовірность помилки візьмемо на рівні $\alpha_а = 0{,}05$, а для ринку оренди комерційної нерухомості (див. рис. 3.1, *б*) – $\alpha_б = 0{,}01$.

Для обох випадків нуль-гіпотезу формулюємо як H_0:$R(x,y)=0$. Альтернативна гіпотеза – двобічна: H_1:R(x,y) $\neq 0$, припускаємо наявність зв'язку, але його напрям не уточнюється.

Вибіркове значення *t*-критерію за формулою (3.6) становитиме:

– для продуктивності праці (див. рис. 3.1, *а*):

$$t_а = \sqrt{\frac{0{,}90^2 \bullet (8-2)}{1-0{,}90^2}} = \sqrt{\frac{0{,}81 \bullet 6}{1-0{,}81}} = \sqrt{\frac{4{,}86}{0{,}19}} = 25.58;$$

– для ринку оренди комерційної нерухомості (рис. 3.1, *б*):

$$t_б = \sqrt{\frac{(-0{,}987)^2 \bullet (7-2)}{1-(-0{,}987)^2}} = \sqrt{\frac{0{,}974 \bullet 5}{1-0{,}974}} = \sqrt{\frac{4{,}87}{0{,}026}} = \sqrt{188{,}538} = 13{,}66.$$

Табличне значення *t**-критерію за формулою (3.6) можна визначити за допомогою статистичних таблиць (див. розділ 2). Використовуючи сучасні версії програм Excel, зокрема англомовні онлайн-застосунки, розрахунок двобічного критичного значення *t**-критерію значно спрощується: слід використати функцію T.INV.2T, а в її параметрах імовірність помилки α ділити на 2 не потрібно. Таким чином, критичні значення *t** є такими:

– для продуктивності праці (див. рис. 3.1, *а*):

$$t^*_а\left(\frac{0{,}05}{2};8-2\right) = t^*_а\left(0{,}025;6\right) = 2{,}447,$$

в клітинку процесора Excel19 введено формулу: =T.INV.2T(0.05;8-2);

– для ринку оренди комерційної нерухомості (див. рис. 3.1, *б*):

$$25{,}58 > 2{,}447 \Rightarrow \text{H}_0 \Rightarrow \text{H}_1(\text{i}),\ t^*_б\left(\frac{0{,}01}{2};7-2\right) = t^*_а\left(0{,}005;5\right) = 4{,}032,$$

в клітинку процесора Excel19 введено формулу: =T.INV.2T(0.01;7-2).

Рішення верифікації: в обох випадках слід відкинути нуль-гіпотезу про брак зв'язку та прийняти альтернативну про те, що є кореляційний зв'язок між x та y:

– для продуктивності праці (див. рис. 3.1, *а*):

$$25{,}58 > 2{,}447 \Rightarrow \textbf{H0} \Rightarrow \textbf{H1(i);}$$

– для ринку оренди комерційної нерухомості (див. рис. 3.1, *б*):

$$13{,}66 > 4{,}032 \Rightarrow \text{Н}_0 \Rightarrow \text{Н}_1(\text{і});$$

$$t^*_{б}\left(\frac{0{,}01}{2};7-2\right) = t^*_{а}\left(0{,}005;5\right) = 4{,}032\,.$$

3.2. РЕГРЕСІЙНІ ЗАЛЕЖНОСТІ. МНК

3.2.1. Основні теоретичні відомості

Після перевірки наявності та щільності кореляційного зв'язку між ознаками доцільно побудувати регресійне рівняння, яке дасть змогу прогнозувати значення результаційної ознаки на основі відомої факторної. Наприклад, знаючи віддаль торговельних площ від транспортних зупинок, за допомогою регресійного рівняння не складно обґрунтувати найвигіднішу орендну ставку і для орендаря, і для орендодавця. Регресійні залежності не обов'язково мають бути лінійними, наприклад формула на рис. 3.1, *а*, що відображає залежність обсягу робіт від трудовитрат, є нелінійною, адже з мікроекономіки добре відомий закон про спадну граничну продуктивність змінного фактора. Вибір математичної форми регресійного рівняння називають також **специфікацією моделі.** Помилка специфікації спотворює результати моделювання: не завжди фактичні дані характеризуються лінійною залежністю. Наприклад, апроксимуючи залежність між ціною та обсягом попиту, можна припуститись значних помилок на деяких діапазонах обсягу продажу (рис. 3.2.).

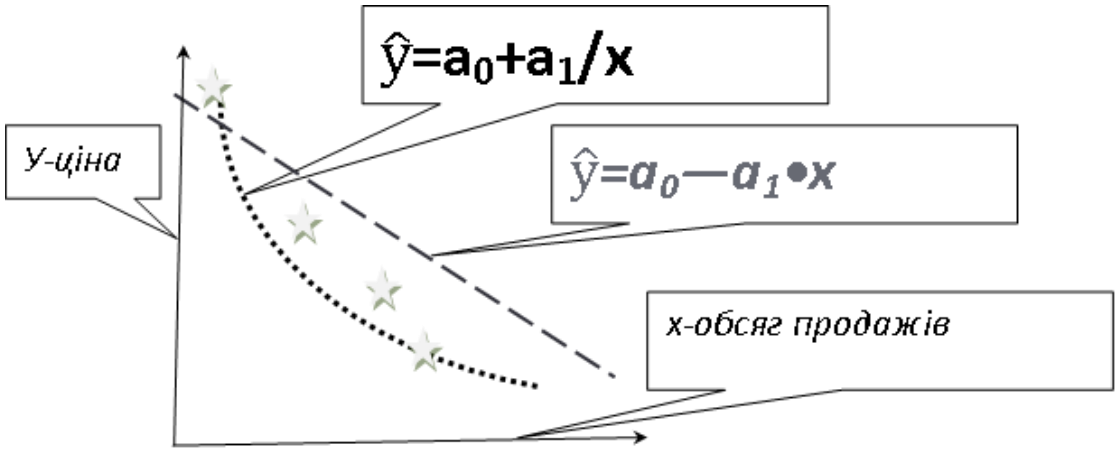


Рис. 3.2. Процес специфікації регресійної моделі та ймовірні помилки

Специфікація – аналітична форма економетричної моделі, що являє собою певну функцію. У рівнянні функції враховано те, що фактичне значення y й теоретичне значення $\hat{y}$, розраховане за моделлю, можуть не збігатися. Різниця між ними – **помилка апроксимації**. Помилку апроксимації позначають як e, це різниця між фактичним і теоретичним значеннями результаційної ознаки:

$$e = y - \hat{y}. \qquad (3.8)$$

Основні типи парних регресійних моделей такі:

- **лінійна:**

  $y = a_0 + a_1 \cdot x + e;$

  $\hat{y} = a_0 + a_1 \cdot x.$

- **нелінійні:**

  – гепрболічна

  $y = a_0 + \frac{a_1}{x} + e;$

  $\hat{y} = a_0 + \frac{a_1}{x};$

  – ступенева

  $y = a_0 \cdot (x)^{a_1} + e;$

  $\hat{y} = a_0 \cdot (x)^{a_1};$

  – параболічна

  $y = a_0 + a_1 \cdot x + a_2 \cdot (x)^2 + e;$

  $\hat{y} = a_0 + a_1 \cdot x + a_2 \cdot (x)^2;$

  – комбінована, поєднує різні типи лінійного та нелінійного зв'язку, наприклад:

  $\hat{y} = a_0 + a_1 \cdot x + \frac{a_2}{x} + x^{a_3}.$

У всіх наведених формулах:

x – незалежна, або факторна, змінна;

y – залежна, результаційна змінна;

a_0, a_1, a_2, a_3 – коефіцієнти (або константи, або параметри) регресійних моделей, які, власне, і визначають за допомогою регресійного аналізу даних вибірки.

Отже, для побудови регресійного рівняння спочатку доводиться розв'язувати обернену задачу: за відомими наборами x та y визначити оцінки коефіцієнтів a_0, a_1, а надалі отримані рівняння використовують вже «традиційно»: замість x підставляються певні значення й за одержаними емпіричними формулами розраховують математичне сподівання y.

Обґрунтовуючи аналітичні рівнянні лінійних та нелінійних регресій, тобто розрахунок оцінок коефіцієнтів (***a*₀**, ***a*₁**, …), найчастіше застосовують **метод найменших квадратів (МНК)**. Його алгоритм реалізовано в більшості вбудованих функцій комп'ютерних програм.

3.2.2. Сутність МНК

МНК потребує визначення параметрів економетричної моделі у такий спосіб, щоб мінімізувати суму квадратів помилки моделі:

$$\sum_{i=1}^{n}(y - \hat{y})^2 \to \min. \qquad (3.9)$$

Для найпростішої парної лінійної залежності формула (3.9) набуває вигляду

$$\sum_{i=1}^{n}(y - \hat{y})^2 = \sum_{i=1}^{n}(y - (a_0 + a_1 \cdot x))^2 \to \min. \qquad (3.10)$$

Графічно МНК мінімізує розсіювання фактичних спостережень навколо теоретичної лінії (рис. 3.3).

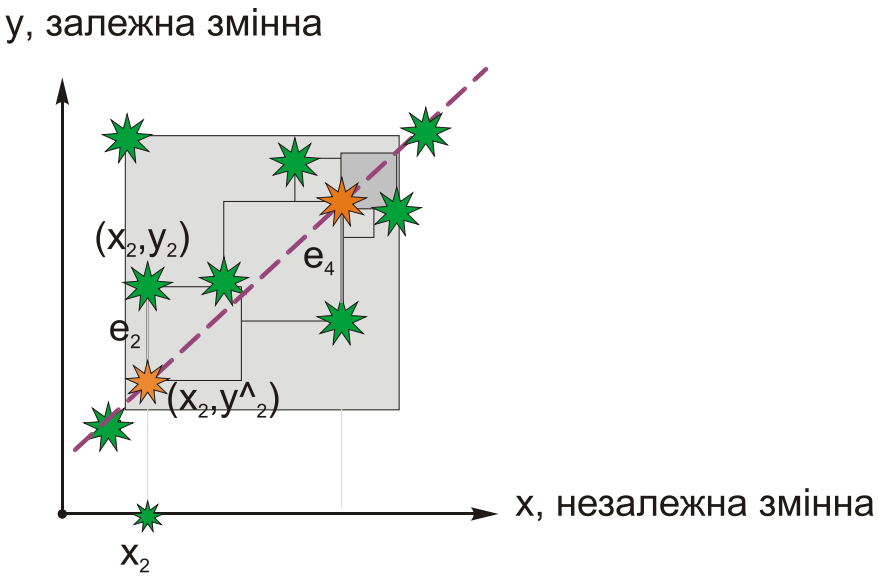


Рис. 3.3. Графічна інтерпретація МНК

Мінімуму суми квадратів помилок досягають за умови рівності нулеві похідної формули (3.9) чи (3.10). Похідну визначають за всіма коефіцієнтами: і a_0, і a_1, і, за потреби, за іншими, як у комбінованій та параболічній моделях. Оскільки доводиться обчислювати декілька похідних, отримані результати диференціювання являють собою систему рівнянь:

$$\begin{cases} \left(\sum_{i=1}^{n} (y - (a_0 + a_1 \cdot x))^2 \right)'_{a_0} = 0; \\ \left(\sum_{i=1}^{n} (y - (a_0 + a_1 \cdot x))^2 \right)'_{a_1} = 0. \end{cases} \quad (3.11)$$

Для подальших викладок наведемо приклади розрахунків похідної простої та складної квадратичної функції:

$$q=4 \cdot a^2 \qquad q'=4 \cdot 2 \cdot a^{2-1}=4 \cdot 2 \cdot a;$$
$$q=((4 \cdot a+3)^2)'=2 \cdot (4 \cdot a+3) \cdot 4.$$

У другому рівнянні сума (4•a+3) являє собою складну функцію *z*, тому спочатку диференціюють квадратичну функцію: залишається сума в дужках і виникає співмножник 2. Далі диференціювання виразу в дужках приводить до появи множника 4. Аналогічна ситуація відбувається і з символьними викладками у разі диференціювання за невідомою поки що величиною a_1. Співмножник *x* у формулі (3.10) – аналог константи 4, тому друге з рівнянь системи буде квадратичним, а не лінійним:

$$\begin{cases} \sum_{i=1}^{n} 2 \cdot \underbrace{(y_i - (a_0 + a_1 \cdot x_i))}_{=0} = 0; \\ \sum 2 \cdot \underbrace{(y_i - (a_0 + a_1 \cdot x_i)) \cdot x_i}_{=0} = 0. \end{cases} \quad (3.12)$$

Рівність нулеві рівнянь (3.12) пояснюється сумою результатів у дужках, тобто за знак сигми можна було винести співмножник 2. Таким чином одержимо іншу еквіваленту систему:

$$\begin{cases} \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (a_0 + a_1 \cdot x_i); \\ \sum_{i=1}^{n} y_i \cdot x_i = \sum_{i=1}^{n} (a_0 + a_1 \cdot x_i) \cdot x_i. \end{cases} \quad (3.13)$$

Розкриваючи дужки у системі рівнянь (3.13) та виконуючи спрощення, отримуємо систему нормальних рівнянь:

$$\begin{cases} \sum y = a_0 \cdot n + a_1 \cdot \sum x; \\ \sum y \cdot x = a_0 \cdot \sum x + a_1 \cdot \sum (x)^2. \end{cases} \quad (3.13)$$

У рівняннях (3.13) немає індексів *i***,** це дало змогу зробити запис більш компактним без втрати змісту.

Одержану систему нормальних рівнянь (3.13) можна розв'язати за допомогою матричних методів (методом Крамера). Втім, використовуючи метод Гауса, а також правила обчислення середніх і всі відомі з попередніх

розділів формули «сирого розрахунку», можна встановити зв'язки між різними статистичними показниками**:**

$$\begin{cases}\sum y = a_0 \bullet n + a_1 \bullet \sum x \\ \sum y \bullet x = a_0 \bullet \sum x + a_1 \bullet \sum x^2\end{cases} \Rightarrow a_0 = \frac{\sum y - a_1 \bullet \sum x}{n} = \underset{\downarrow \bar{y}}{\frac{\sum y}{n}} - a_1 \bullet \underset{\downarrow \bar{x}}{\frac{\sum x}{n}};$$

$$a_0 = \bar{y} - a_1 \bullet \bar{x}.$$

Для того щоби знайти a_1 треба змінити систему так, щоб у ній не було a_0:

$$\begin{cases}\sum y = a_0 \bullet n + a_1 \bullet \sum x \ \Big| \bullet \sum x \\ \sum y \bullet x = a_0 \bullet \sum x + a_1 \bullet \sum x^2 \ \Big| \bullet n\end{cases} \Rightarrow \begin{cases}\sum y \bullet \sum x = a_0 \bullet n \bullet \sum x + a_1 \bullet \sum x \bullet \sum x \\ \sum y \bullet x \bullet n = a_0 \bullet \sum x \bullet n + a_1 \bullet \sum x^2 \bullet n.\end{cases}$$

Для розрахунку a_1 другого рівня віднімаємо перше. При цьому зникне складник з a_0:

$$\sum y \bullet x \bullet n - \sum y \bullet \sum x = a_1 \bullet \sum x^2 \bullet n - a_1 \bullet \sum x \bullet \sum x;$$

$$a_1 \bullet \left(\sum x^2 \bullet n - \left(\sum x\right)^2\right) = \sum y \bullet x \bullet n - \sum y \bullet \sum x;$$

$$a_1 = \frac{\left(\sum y \bullet x \bullet n - \sum y \bullet \sum x\right)/n^2}{\left(\sum x^2 \bullet n - \left(\sum x\right)^2\right)/n^2} = \frac{\overline{y \bullet x} - \bar{y} \bullet \bar{x}}{\overline{x^2} - \bar{x}^2} \Rightarrow a_1 = \frac{cov(x,y)}{\sigma^2{}_x}.$$

Для однофакторної регресії формули коефіцієнтів такі:

- a_1 – нахил, показує, як швидко (стрімко чи положисто) зростає або спадає теоретична лінія регресії. Для того щоби його знайти, потрібно коваріацію фактора та результату поділити на дисперсію x:

$$a_1 = \frac{\overline{y \bullet x} - \bar{y} \bullet \bar{x}}{\overline{x^2} - \bar{x}^2} \Rightarrow a_1 = \frac{cov(x,y)}{\sigma^2{}_x}. \qquad (3.14)$$

Цей показник можна розрахувати в Excel за допомогою функції SLOPE (НАКЛОН для ранніх версій**)**;

- a_0 – відрізок, або вільна константа, що показує значення *y* за нульового значення *x*. Для того щоби його обчислити, треба скоригувати середній рівень результативного показника $\bar{y}$ на середній рівень фактора $\bar{x}$, помножений на коефіцієнт нахилу a_1:

$$a_0 = \bar{y} - a_1 \bullet \bar{x}. \qquad (3.15)$$

3.2.3. Умови застосування МНК

У випадку парної регресійної залежності МНК потребує п'ятьох гіпотез. Якщо будують складніші багатофакторні моделі, кількість гіпотез зростає до шістьох.

Основні гіпотези, потрібні для побудови парних регресій:

1. *Є лінійний зв'язок* між результаційною та факторними змінними, який характеризується рівнянням регресії: $y=a_0+a_1 \cdot x+e$, $\hat{y}=a_0+a_1 \cdot x$. У разі нелінійного зв'язку *y* та *x*-и необхідно *лінеаризувати* (звести до лінійного вигляду).
2. Факторна змінна *x є невипадковою (детерміністичною) величиною*, тобто це – «керована» змінна, впливаючи на її величину, можна змінити *y* (залежну змінну).
3. *Математичне сподівання (середнє значення) випадкового вектора помилок дорівнює нулю, а дисперсія e (помилки) є невеликою постійною додатною величиною*, яка не залежить від індекса *i*. Умова незалежності величини помилки від номера спостереження:

$$M(e_i)=0; \qquad (3.16, а)$$

$$D(e_i)=M((e_i)\text{^}2)=\sigma^2. \qquad (3.16, б)$$

Якщо

- умова незалежності *x* та *e* порушена, $R(x,e)>0,5$, виникає явище, що має назву «гетероскедастичність», а отримана модель може виявитись хибною;
- *x* та *e* незалежні, кореляційного зв'язку між ними немає, спостерігається гомоскедастичність моделі, певніше за все така модель є правільною, тоді теоретичні значення $\hat{y}$, розраховані за її допомогою, будуть максимально наближатись до фактичних значень *y*.

У випадку гетероскедастичності помилка моделі залежить функціонально від незалежної змінної, тому дисперсія помилки моделі, а також надійний інтервал для моделі будуть оцінені неточно: найпевніше надійний інтервал буде дуже широким.

Для перевірки на гетероскедастичність можна або розрахувати коефіцієнт кореляції між *x* та *e*, або побудувати модель, у якій *e* буде залежати від *x.* Така модель може бути як лінійною, так і нелінійною. Рішення про існування гетероскедастичності може бути ухвалене на основі коефіцієнта детермінації вторинної регресійної моделі для *e* як функції від *x* (коефіцієнт достовірності апроксимації моделі *e* від *x*).

4. *Компоненти вектора E* (набір помилок у всіх спостереженнях) є некорельованими величинами. Порушення цієї гіпотези має назву – *автокореляція залишків моделі.* Вона трапляється під час аналізу часових рядів. Для автокореляції характерна повторюваність значень залишків для різних спостережень.
5. Часто вважають, що *помилка має нормальний закон розподілу з нульовим математичним сподіванням та невеликою дисперсією*. Якщо ця гіпотеза є істинною, то лінійна регресія – *класичною* $e \sim N(0,\sigma^2)$.

6. *Факторні ознаки* лінійного множинного рівняння *не корельовані між собою* $\boxed{\mathrm{Cor}(x_j,\ x_k) \sim 0}$. *Порушення гіпотези свідчить* про *мультиколінеарність*. У випадку мультиколінеарності *зростає не лише дисперсія помилки моделі e*, а й дисперсія помилки *коефіцієнтів моделі* a_j. Тоді коефіцієнти моделі мають завеликий надійний інтервал, а сама модель втрачає точність.

3.2.4. Показники якості регресійної моделі

Для визначення якості (точності) регресійної моделі розраховують такі показники.

- **Дисперсія помилки:**

$$S^2_{e_{виправл}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2} = \sigma_e^2 . \qquad (3.17)$$

За формулою (3.17) визначають незміщену, або виправлену, дисперсію. У знаменнику цієї формули – кількість ступенів вільності регресійної моделі. Оскільки в розрахунках коефіцієнтів використовують два середніх: і для факторної ознаки ($\bar{x}$), і для результаційною ($\bar{x}$), від n віднімають 2: $\nu = n - 1 - 1 \Rightarrow \nu = n - 2$**.** Аналогічно визначають ступені вільності для коефіцієнта парної кореляції у п. 3.1.3.

Та якщо в знаменнику використати не кількість ступенів вільності, а обсяг вибірки, одержимо зміщену оцінку дисперсії помилки:

$$S_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n} .$$

- Стандартна помилка моделі (стандартне відхилення помилки):

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}} . \qquad (3.18)$$

Формула (3.18) дає змогу визначити виправлену, незміщену помилку, тоді як використання під коренем знаменника обсягу вибірки призводить до обчислення зміщеної, невиправленої помилки:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}} . \qquad (3.19)$$

Звичайно, результат (3.18) буде вищим, а розрахована за виразом (3.19) величина s_e дещо зменшує помилку регресії. Втім, на великих вибірках, коли n>30, різниця між σ_e та s_e буде досить незначною.

- Коефіцієнт детермінації (достовірності апроксимації):

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}. \qquad (3.20)$$

Саме цю величину наведено на рис. 3.1, *а*, *б* для вимірювання точності наближення теоретичної регресійної лінії до хмари точок фактичних даних. Звісно, беручи до уваги раніше наведені формули, вираз (3.20) можна деталізувати:

$$R^2 = 1 - \frac{\dfrac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}}{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}. \qquad (3.21)$$

Коли модель є точною, помилка набуває малих, близьких до 0 значень. Тоді наближається до нуля і дисперсія помилки, і співвідношення у формулі R^2. Тобто R^2 наближається до 1. Що ближче R^2 до 1, то точнішою є модель:

$$e = y - \hat{y} \approx 0 \to \sigma_e^2 \approx 0 \to \frac{\sigma_e^2}{\sigma_y^2} \approx 0 \to R^2 \approx 1.$$

Для визначення щільності зв'язку застосовують шкалу Чеддока (табл.3.3):

Таблиця 3.3

Шкала Чеддока для оцінювання достовірності апроксимації регресійних моделей

Значення R^2	**0,1 — 0,3**	**0,3 — 0,5**	**0,5 — 0,7**	**0,7 — 0,9**	**0,9 — 0,99**
Оцінка щільності зв'язку	слабкий	помірний	помітний (середній)	значний (сильний)	дуже значний (дуже сильний)

- Критерій Фішера (F-критерій) показує, чи не має модель випадкового характеру. Він являє собою співвідношення частини дисперсії у (σ_y^2), пояснюваної за допомогою парної регресії (R^2), та частини дисперсії у, регресійної моделі якої не можна пояснити ($1-R^2$):

$$F = \frac{R^2}{1 - R^2}. \qquad (3.22)$$

Використання для обчислень F-критерію показника достовірності апроксимації пояснюється такими співвідношеннями:

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2} = \frac{\sigma_y^2}{\sigma_y^2} - \frac{\sigma_e^2}{\sigma_y^2} = \frac{\sigma_y^2 - \sigma_e^2}{\sigma_y^2} = \frac{\text{пояснена дисперсія (мінливість)}\, y}{\text{загальна дисперсія}\, y}.$$

Отримане значення F-критерію для моделі порівнюють з критичним F*, обчисленим за певної довірчої імовірності та кількості ступенів вільності $\nu_1 = 1$ (оскільки у моделі тільки одна незалежна змінна –, x) й $\nu_2 = n - 2$ (таке саме як й у решті випадків кількості ступенів вільності для дослідження зв'язку між двома ознаками). Верифікація нуль-гіпотези про випадковий характер зв'язку між x та y (H_0) потребує порівняння F-моделі та $F^*_{\alpha;1;n-2}$.

Якщо $F > F^*_{\alpha,1,n-2} \Rightarrow$ ~~H_0~~→H_1 (і) →, залежність між *x* та *y* є невипадковою.

- Стандартна помилка коефіцієнта моделі:
  - для коефіцієнта нахилу регресії:

$$\sigma_{a_1} = \sqrt{\frac{\sigma_e^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}; \tag{3.23}$$

  - для вільної константи:

$$\sigma_{a_0} = \sqrt{\frac{\sigma_e^2 \cdot \sum_{i=1}^{n} x_i^{\,2}}{n \cdot \sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}}. \tag{3.24}$$

- Півширина надійного інтервалу коефіцієнта моделі:

$$\Delta_{a_j} = t^*_{\frac{\alpha}{2};n-2} \cdot \sigma_{a_j}, \tag{3.25}$$

де *j* – порядковий номер регресійного коефіцієнта. Для парної регресії *j*=0 (вільна константа), *j*=1 (коефіцієнт при змінній *x*).

- Перевірка значущості коефіцієнта моделі *t*-критерій коефіцієнта (t_{aj})

Значення окремих коефіцієнтів моделі (a_0,a_1) перевіряють за *t*-критерієм, який має бути вищим за табличне значення *t**. Розраховують *t*-критерій для кожного коефіцієнта, зокрема вільної константи, беручи до уваги помилку коефіцієнта моделі:

$$t_j = \left|\frac{a_j}{\sigma_{a_j}}\right| \rightarrow t_j < t^*_{\left(\frac{\alpha}{2};\nu=n-2\right)} \rightarrow \text{фактор малозначущий.} \tag{3.26}$$

Прямі дужки означають, що для аналізу важливим є модуль співвідношення коефіцієнта та його стандартної помилки, оскільки коефіцієнти регресії, і, відповідно, їхні співвідношення із стандартною помилкою, можуть набувати від'ємних значень.

- Перевірка зміщеності коефіцієнта моделі:

$$\left|\frac{a_j}{\sigma_{a_j}}\right| - 1 < 0{,}1. \tag{3.27}$$

Якщо нерівність є правильною, коефіцієнт вважають зміщеним, що негативно позначається на якості й точності моделі.

$$e_i \uparrow \rightarrow \sigma_e \uparrow \rightarrow \sigma_{a_j} \uparrow \rightarrow \begin{cases} \rightarrow \Delta a_j \uparrow \\ \rightarrow = \dfrac{a_j}{\sigma_{a_j}} = t_j \rightarrow t_j < t^*_{\frac{\alpha}{2};n-2} \rightarrow \text{фактор малозначущий} \end{cases}$$

Зміщеність коефіцієнта моделі полягає в тому, що помилки для спостережень у переважній більшості будуть завищеними, або ж навпаки, заниженими порівняно з фактичними значеннями *y*.

- Півширина надійного інтервалу прогнозу моделі ($\Delta\hat{y}$):

$$\Delta\hat{y} = t^*_{\frac{\alpha}{2};n-2} \cdot \sigma_e \cdot \sqrt{\frac{1}{n} + \frac{\left(\hat{x} - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}, \qquad (3.28)$$

де $\hat{x}$ – значення факторної ознаки, для якого розраховують величину результативної ознаки $\hat{y}$.

Надійний інтервал прогнозу моделі визначають подібно до інших формул обрахунку півширини надійного інтервалу малих вибірок, тобто використовують табличні (двобічні) значення критерію Стьюдента (*t**), а також стандартну помилку моделі (σ_e). Втім, у формулі (3.28) використано поправний множник, який означає, що прогнозне значення $\hat{x}$, яке підставляють у модель, може бути більшим або меншим, аніж максимальне чи мінімальне значення x вихідної вибірки, тобто $\hat{x} \notin [x_{\min}; x_{\max}]$.

Залежність півширини надійного інтервалу прогнозу $\Delta\hat{y}$ від прогнозного значення $\hat{x}$ (3.28) має квадратичний характер, допустимі межі відхилення фактичних значень y від теоретичних оцінок $\hat{y}$ мають вигляд парабол, які зверху й знизу охоплюють теоретичну лінію регресії. Вершини парабол, отже, і найвужчі надійні інтервали є відповідними середньому значенню факторної ознаки ($\bar{x}$). У міру віддалення в обидва боки $\hat{x}$ від $\bar{x}$ розширюються й межі надійного інтервалу помилки. Тому прогностичні розрахунки за регресійними рівняннями у разі значного віддалення $\hat{x}$ від x_{max} і від $x_{мін}$ можуть виявитись малозначущими.

> ***Приклад 3.4.*** За даними рис. 3.1, *б* та прикладів 3.1. – 3.3 побудуйте парну лінійну регресію залежності ставки оренди (*y*) від відстані між об'єктом і зупинкою громадського транспорту (*x*). Проаналізувати показники якості побудованої моделі.

Параметри парного регресійного рівняння можна визначити за формулами (3.14) та (3.15), оскільки раніше вже обчислено коваріацію та середні значеннях та їхніх квадратів:

$$\text{cov}(x,y)=-9{,}01;\ \bar{x}=5{,}04;\ \bar{y}=5{,}16;\ \overline{x^2}=34{,}69.$$

- a_1 – нахил регресії, показує, наскільки змінюється ставка оренди у міру збільшення віддалі об'єкта від зупинки громадського транспорту:

$$a_1=\frac{\overline{y\bullet x}-\bar{y}\bullet\bar{x}}{\overline{x^2}-\bar{x}^2}=\frac{-9{,}01}{34{,}69-5{,}04^2}=-0{,}97 \text{ тис. грн/км}.$$

- a_0 – вільна константа, відображає вплив на вартість оренди інших факторів, не включених до моделі (інфляція, курс валюти, рівень цін на комунальні послуги та ін.):

$$a_0=\bar{y}-a_1\bullet\bar{x}=5{,}16-(-0{,}97)\bullet 5{,}04=5{,}16+4{,}89=10{,}05 \text{ тис.грн.}$$

Таким чином, отримано емпіричну формулу для визначення вартості оренди 1 м^2 торговельних площ строком один місяць:

$$\hat{y}=10{,}05-0{,}97\bullet x, \qquad (3.29)$$

де x – відстань від найближчої зупинки громадського транспорту, км;

y – ставка орендної плати, грн•м^2/міс.

Параметри рівняння (3.29) майже збігаються із записом формули на рис.3.1, *б*, незначна розбіжність у значенні вільної константи (10,05 проти 10,07) пов'язана із округленнями, оскільки Excel заокруглює проміжні обчислення до 10 десяткових знаків, у той час, як у посібнику заокруглення обмежено лише трьома десятковими знаками.

Аналогічну виразу (3.29) модель можна було б отримати, склавши систему нормальних рівнянь:

$$\begin{cases}\sum y=a_0\bullet n+a_1\bullet\sum x\\ \sum y\bullet x=a_0\bullet\sum x+a_1\bullet\sum(x)^2\end{cases}\Rightarrow\begin{cases}36{,}1=a_0\bullet 7+a_1\bullet 35{,}3\\ 118{,}95=a_0\bullet 35{,}3+a_1\bullet 242{,}8\end{cases}\Rightarrow\begin{cases}a_1=-0{,}974\\ a_0=10{,}068\end{cases}$$

$$\Downarrow$$

$$\hat{y}=10{,}068-0{,}974\bullet x. \qquad (3.29)$$

Рівняння (3.29) більшою мірою збігається із рівнянням на рис. 3.1,*б*, оскільки обчислення виконано над первинними, незаокругленими даними.

Для побудови системи нормальних рівнянь обчислено сумарні значення кожної з ознак та їхніх добутків:

$$\bar{x}=\frac{35{,}3}{7}=5{,}04\Rightarrow\sum_{i=1}^{n}x_i=35{,}3;\qquad \bar{y}=\frac{36{,}1}{7}=5{,}16\Rightarrow\sum_{i=1}^{n}y_i=36{,}1;$$

$$\overline{x^2}=\frac{242{,}8}{7}=34{,}69\Rightarrow\sum_{i=1}^{n}x_i^2=242{,}8;$$

$$\overline{x\bullet y}=\frac{118{,}95}{7}\approx 17\Rightarrow\sum_{i=1}^{n}x_i\bullet y_i=118{,}95.$$

Для розрахунку низки показників, які характеризують якість й точність побудованої моделі, потрібно скласти таблицю допоміжних розрахунків (табл. 3.4).

Таблиця 3.4

Допоміжні розрахунки для оцінювання якості регресійного рівняння

№ спостереження	Фактична ставка оренди, грн•м²/міс., y_i	Теоретична ставка оренди, грн•м²/міс., згідно з виразом (3.29), $\hat{y}_i$, $\hat{y}_i = 10{,}068 - 0{,}974 \cdot x_i$	Помилка моделі, e^2, грн•м²/міс., $e = y_i - \hat{y}_i$	Квадрат помилки моделі, e^2, грн•м²/міс., $e^2 = (y_i - \hat{y}_i)^2$	Відхилення від середнього результаційної ознаки $y_i - \bar{y} = y_i - 5{,}16$	Квадрат відхилення від середнього результаційної ознаки $(y_i - \bar{y})^2 = (y_i - 5{,}16)^2$
1	2	3	4	5	6	7
1	10,40	9,19=10,068-0,974•0,9	1,21 =10,4-9,19	1,4641	5.24	27.4576
2	8,20	8,41=10,068-0,974•1,7	-0,21 =8,2-8,41	0,0441	3.04	9.2416
3	6,40	6,66=10,068-0,974•3,5	-0,26 =6,4-6,66	0,0676	1.24	1.5376
4	4,00	5,49=10,068-0,974•4,7	-1,49 =4-5,49	2,2201	-1.16	1.3456
5	3,80	3,74=10,068-0,974•6,5	0,06 =3,8-3,74	0,0036	-1.36	1.8496

Закінчення табл. 3.4

1	2	3	4	5	6	7
6	1,70	1,79=10,068-0,974•8,5	-0,09 =1,7-1,79	0,0081	-3.46	11.9716
7	1,60	0,82=10,068-0,974•9,5	0,78 =1,6-1,79	0,6084	-3.56	12.6736
Разом	36,10	36,1	0	4,42	**0**	66,0772

За підсумковим рядком табл. 3.4 визначаємо проміжні показники для оцінювання якості моделі за формулою (3.29):

– сума квадратів помилки регресії $\sum_{i=1}^{n}(y_i - \hat{y})^2 = 4{,}61$;

– сума квадратів результаційної ознаки $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 66{,}08$;

– дисперсія результаційної ознаки $\sigma_y^2 = \frac{\sum_{i=1}^{n}(y - \bar{y})^2}{n-1} = \frac{66{,}08}{7-1} = \frac{66{,}08}{6} = 11{,}013$.

Таким чином, модель (3.9) характеризують

- виправлена дисперсія помилки:

$$\sigma_e^2 = \frac{\sum_{i=1}^{n}(y - \hat{y})^2}{n-2} = \frac{4{,}42}{8-2} = \frac{4{,}42}{6} = 0{,}736 .$$

При цьому зміщена оцінка дисперсії помилки є дещо нижчою:

$$s_e^2 = \frac{\sum_{i=1}^{n}(y - \hat{y})^2}{n} = \frac{4{,}42}{7=} = 0{,}631;$$

- виправлена стандартна помилка моделі (стандартне відхилення помилки):

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n}(y - \hat{y})^2}{n-2}} = \sqrt{\sigma_e^2} = \sqrt{0,736} = 0,858 \left(\text{тис. грн•м}^2/\text{міс}\right).$$

Невиправлена помилка є дещо меншою:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y - \hat{y})^2}{n}} = \sqrt{s_e^2} = \sqrt{\frac{4,42}{7}} = \sqrt{0,631} = 0,812 < 0,783 ;$$

- коефіцієнт детермінації (достовірності апроксимації):

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2} = 1 - \frac{0,858}{11,013} = 0,922 .$$

Результат виявився дуже близьким до R^2 на рис. 3.1, *б*, деякі розбіжності на рівні сотих пов'язані з округленнями: 0,922 проти 0,9301. Можна стверджувати, що модель (3.29) пояснила варіацію орендних ставок внаслідок різного розміщення об'єктів вибірки на 92,2%, і лише 7,8%(=(1—,922)•100%) варіації *y* залишилось непоясненою. Тобто модель характеризується дуже сильною оцінкою щільності зв'язку за шкалою Чеддока:

$$0,922 \in [0,9 - 0,99).$$

Можна стверджувати, що між фактором розміщення (відстань від зупинки транспорту *x*) та місячною орендною ставкою торговельних приміщень (*y*) є зв'язок, наближений до лінійного, оскільки $0,922\left(= R^2\right) \to 1$. Цей висновок ілюструє графік (рис.3.19, б), більшість точок спостереження на якому «тяжіє» до регресійної прямої.

Однак потрібно виконати додаткову перевірку, чи не є виявлена залежність випадковою – результатом «вдало» зібраної малої вибірки. Для цього використовують F-критерій, який дуже часто спростовує позитивні висновки, зроблені на основі коефіцієнта детермінації R^2;

- критерій Фішера (F-критерій) моделі (3.29):

$$F = \frac{R^2}{1 - R^2} = \frac{0,922}{1 - 0,922} = \frac{0,922}{0,078} = 11,82 .$$

Критичне F* обчислимо за довірчої імовірності 0,05 та кількості ступенів вільності $\nu_1 = 1$ (оскільки у моделі тільки одна незалежна змінна, x) й $\nu_2 = n - 2 = 7 - 2 = 5$ (так саме, як й у решті випадків кількості ступенів вільності для вивчення зв'язку між двома ознаками). Критичне значення: $F^*_{0,05;1;5} = 6,608$, його визначено за допомогою функції Excel:

=F.INV.RT(0.05;1;5), аналогічний результат можна було б отримати за допомогою таблиць попереднього розділу.

Оскільки $F > F^*_{0,05;1;5}\left(11{,}82 > 6{,}608\right) \Rightarrow \text{H}_0 \rightarrow \text{H}_1$ (і) у 95 випадках з 100 залежність (3.29) між x та y є невипадковою.

Однак, обираючи надійну імовірність на рівні 0,01, як й у прикладі 3.3, одержимо значно більше критичне значення F*, яке підтвердить справедливість нуль-гіпотези про випадковий характер моделі (3.29):

$F^*_{0,01;1;5} = 16{,}258$, його визначено за допомогою функції Excel:

=F.INV.RT(0.01;1;5).

Тоді $F < F^*_{0,01;1;5}\left(11{,}82 < 16{,}258\right) \Rightarrow \text{H}_0$ (і), отже, не можна стверджувати, що залежність (3.29) між x та y виявиться невипадковою й у 99 випадках зі 100.

Звісно, довірча імовірність 0,01 є досить високою для економічних моделей, а вибірка із сімох об'єктів – дуже малою, тому остаточно ухвалюємо висновок про невипадковий характер регресійної формули (3.29), зроблений за імовірності помилки $\alpha = 0{,}05$.

Стандартну помилку коефіцієнтів моделі визначають із урахуванням суми квадратів відхилень від середнього показників розміщення об'єктів. Цю суму квадратів можна визначити на основі невиправленої дисперсії x, визначену в прикладі 3.2. в розрахунку коефіцієнта кореляції:

$$\sigma_x = \sqrt{\left(34{,}69 - 5{,}04^2\right)} = \sqrt{9{,}29} \Rightarrow \sigma_x^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n} = 9{,}29 \Rightarrow \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 =$$

$$= \sigma_x^2 \bullet n = 9{,}29 \bullet 7 \approx 65{,}03;$$

– для коефіцієнта нахилу регресії:

$$\sigma_{a_1} = \sqrt{\frac{\sigma_e^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}} = \sqrt{\frac{0{,}858}{65{,}03}} = \sqrt{0{,}0135} = 0{,}115 \text{ тис.грн/км};$$

– для вільної константи

$$\sigma_{a_0} = \sqrt{\frac{\sigma_e^2 \bullet \sum_{i=1}^{n} x_i^{\,2}}{n \bullet \sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}} = \sqrt{\frac{0{,}858 \bullet 242{,}8}{7 \bullet 242{,}8 - 35{,}3^2}} = 0{,}678 \text{ тис.грн.}$$

Показники стандартної помилки коефіцієнтів мають таку саму розмірність, що й коефіцієнти регресії.

- Півширину надійного інтервалу коефіцієнтів моделі визначимо на рівні значущості $\alpha = 0{,}05$. При цьому значення коефіцієнта довіри (табличне значення критерію Стьюдента, t*) дорівнюватиме: $t^*_{\frac{\alpha}{2};n-2} = t^*_{\frac{0{,}05}{2};7-2} = 2{,}571$.

Його можна визначити не лише з довідкових статистичних таблиць, а й за допомогою функції Excel: =T.INV.2T(0.05;7-2):

– півширину надійного інтервалу коефіцієнта нахилу регресії, яку визначаємо за виразом (3.25), і яка має таку саму розмірність, що й коефіцієнт a_1:

$$\Delta_{a_1} = t^*_{\frac{0,05}{2};7-2} \cdot \sigma_{a1} = 2,571 \cdot 0,115 = 0,296 \text{ тис.грн/м}^2;$$

– півширину надійного інтервалу для вільної константи також визначаємо за формулою (3.25), її розмірність збігається з коефіцієнтом a_0 й вільною константою:

$$\Delta_{a_0} = t^*_{\frac{0,05}{2};5-2} \cdot \sigma_{a0} = 2,571 \cdot 0,678 = 1,743 \text{ тис.грн}.$$

Таким чином, регресійне рівняння (3.29) можна записати, не позначаючи помилки, але використовуючи надійні інтервали констант:

$$y = 10,068 \pm 1,743 - (0,974 \pm 0,296) \cdot x. \quad (3.30)$$

При цьому в лівій частині рівняння використовуємо символ результаційної ознаки, не позначаючи теоретичного значення. Залежність (3.30) можна записати ще й так:

$$y = 8,325 \ldots 11,811 - (0,678 \ldots 1,266) \cdot x. \quad (3.31)$$

Економічний зміст моделей (3.30) та (3.31) полягає у такому.

- Фактор розміщення, визначуваний відстанню від об'єкта торговельної нерухомості до найближчої зупинки громадського транспорту, має зворотний вплив на ставку оренди. Кожен кілометр віддалення об'єкта знижує вартість винаймання 1 м^2 на місяць в межах від 678 до 1266 грн, або на 0,974 $\pm$ 0,296 тис. грн згідно з коефіцієнтом нахилу a_1;
- Вплив сукупності інших факторів, крім розміщення, не включених до моделі (3.30) чи (3.31), таких як технічний стан торговельних приміщень, поверховість об'єкта оренди, вартість комунальних послуг, чисельність населення й економічна ситуація в районі та інших, додатково підвищує ставку оренди до 8 325 — 11 811 тис. грн, або до 10,068 $\pm$ 1,743 відповідно до вільної константи a_0.

Перевірка значущості коефіцієнтів моделі (3.29):

– коефіцієнта нахилу регресії:

$$t_1 = \left|\frac{a_1}{\sigma_{a_1}}\right| = \left|\frac{-0,974}{0,115}\right| \approx 8,47 \Rightarrow 8,47 > 2,571 \rightarrow t_1 < t^*_{\left(\frac{0,05}{2};5\right)} \rightarrow \text{фактор значущий};$$

– вільної константи:

$$t_0 = \left|\frac{a_0}{\sigma_{a_0}}\right| = \left|\frac{10,068}{0,678}\right| \approx 14,85 \Rightarrow 14,85 > 2,571 \rightarrow t_0 < t^*_{\left(\frac{0,05}{2};5\right)} \rightarrow \text{фактор значущий}.$$

Перевірка зміщеності коефіцієнтів моделі:

– коефіцієнта нахилу регресії:

$$\left|\frac{a_1}{\sigma_{a_1}}\right| - 1 = \left|\frac{-0{,}974}{0{,}115}\right| - 1 \approx 8{,}47 - 1 > 0{,}1 \Rightarrow \text{фактор незміщений};$$

– вільної константи:

$$\left|\frac{a_0}{\sigma_{a_0}}\right| - 1 = \left|\frac{10{,}068}{0{,}678}\right| - 1 \approx 14{,}85 - 1 > 0{,}1 \Rightarrow \text{фактор незміщений}.$$

Зміщеність коефіцієнта моделі полягає в тому, що помилки спостережень будуть переважно завищеними, або ж навпаки, заниженими порівняно із фактичними значеннями ***y***.

Оскільки обидва коефіцієнти моделі (3.29) є значущими та незміщеними, а сама модель характеризується дуже високою достовірністю апроксимації та невипадковим характером залежності, вона є якісною та придатною для прогнозування.

Приклад 3.5. За даними рис. 3.1,*б* та прикладів 3.1 – 3.4 скласти надійний інтервал теоретичної лінії регресії (3.29) й спрогнозувати можливий діапазон орендних ставок для магазину, будівля якого розміщена на відстані 2 км від зупинок громадського транспорту.

На рівні імовірності помилки $\alpha = 0{,}05$ й коефіцієнта довіри $t^*_{\frac{\alpha}{2};n-2} = t^*_{\frac{0{,}05}{2};7-2} = 2{,}571$ півширину надійного інтервалу (3.28) доцільно представити у вигляді залежності від x_i. При цьому, за даними попереднього прикладу: $\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$, $\sigma_e = 0{,}736$ (тис.грн•м^2/міс.), за даними прикладу 3.1 $\bar{x} = 5{,}04$ км і вираз (3.28) запишемо так:

$$\Delta\hat{y} = t^*_{\frac{\alpha}{2};n-2} \bullet \sigma_e \bullet \sqrt{\frac{1}{n} + \frac{\left(\hat{x} - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} = 2{,}571 \bullet 0{,}736 \bullet \sqrt{\frac{1}{7} + \frac{\left(\hat{x} - 5{,}04\right)^2}{65{,}03}}$$

або після деяких обчислень:

$$\Delta\hat{y} = 1{,}892 \bullet \sqrt{\frac{1}{7} + \frac{\left(\hat{x} - 5{,}04\right)^2}{65{,}03}}, \qquad (3.32)$$

Для графічного зображення надійних інтервалів регресійної лінії (рис. 3.4) спочатку складаємо таблицю допоміжних розрахунків (табл.3.5) за допомогою залежності (3.23).

Таблиця 3.5

Розрахунок надійних інтервалів регресійного рівняння (3.29)

Порядковий номер спостереження	Відстань об'єкта від зупинки транспорту, км, x_i	Фактична ставка оренди, грн•м^2/міс., y_i	Теоретична ставка оренди, грн•м^2/міс., згідно (3.29), $\hat{y}_i$, $\hat{y}_i = 10{,}068 - 0{,}974 \cdot x_i$	Півширина надійного інтервалу, грн•м^2/міс., згідно з (3.32) $\Delta\hat{y}$, $\Delta\hat{y} = 1{,}892 \cdot \sqrt{\frac{1}{7} + \frac{(\hat{x} - 5{,}04)^2}{65{,}03}}$	Ліва (нижня) межа надійного інтервалу, грн•м^2/міс., $\hat{y}_i - \Delta\hat{y}_i$	Права (верхня) межа надійного інтервалу, грн•м^2/міс., $\hat{y}_i + \Delta\hat{y}_i$
1	**2**	**3**	**4**	**5**	**6**	
1	0.9	10,40	9,19 =10,068-0,974•0,9	1.21	7.98=9.19-1.21	10.4=9.19+1.21
2	1.7	8,20	8,41 =10,068-0,974•1,7	1.06	7.35=8.41-1.06	9.47=8.41+1.06
3	3.5	6,40	6,66 =10,068-0,974•3,5	0.8	5.86=6.66-0.8	7.46=6.66+0.8
4	4.7	4,00	5,49 =10,068-0,974•4,7	0.72	4.77=5.49-0.72	6.21=5.49+0.72
5	6.5	3,80	3,74 =10,068-0,974•6,5	0.79	2.95=3.74-0.79	4.53=3.74+0.79
6	8,5	1,70	1,79 =10,068-0,974•8,5	1,08	0,71=1,79-1,08	2,87=1,79+1,08
7	9,5	1,60	0,82 =10,068-0,974•9,5	1,27	-0,45=0,82-1,27	2,09=0,82+1,27

Як видно з графи 5 табл. 3.5, значення півширини надійного інтервалу прискорено зменшується у міру наближення до середини від верхнього чи нижнього рядків. В останньому рядку ліва межа надійного інтервалу перейшла нульову позначку, призвівши до від'ємних значень орендної ставки, що, звісно, суперечить здоровому глузду.

Це пояснюється завеликим, порівняно із середнім, значенням факторної ознаки – віддаль від зупинки 9,5 км об'єкта №7 (x_7=9,5) майже вдвічі більша за середнє значення x у вибірці ($\bar{x} = 5{,}04$).

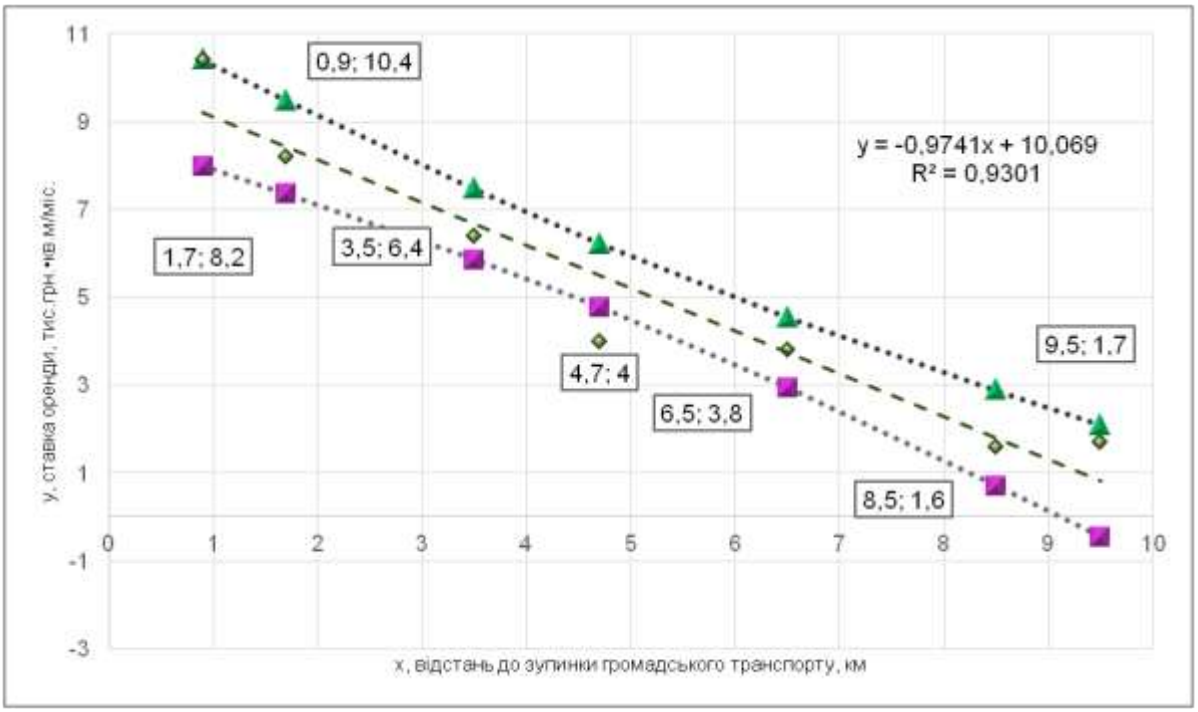


Рис. 3.4. Надійний інтервал лінійної регресії

Конфігурація надійних інтервалів лінійної регресії — криволінійна (рис. 3.4.). Штрихові межі надійних інтервалів найвужчі для об'єкта, розміщеного на відстані 4,7 км від зупинки транспорту. Для об'єктів, розташованих на інших відстанях, як більших, так і менших, надійний інтервал ширший по обидва боки від теоретичної лінії.

На рис. 3.4. одне із спостережень опинилося за межами надійного інтервалу. Це пояснюється тим, що для обґрунтування півширини інтервалу допущено деяку імовірність помилки. Отже, $\alpha = 0{,}05$ у п'ятьох випадках зі 100 орендна ставка може бути як більшою, так і меншою, аніж це передбачене діапазоном теоретичних значень.

Якщо деякий об'єкт розміщено на відстані 2 км ($\hat{x}_i = 2$) від зупинки громадського транспорту, теоретичне значення орендної ставки за виразом (3.29) становитиме:

$$\hat{y}_i = 10{,}068 - 0{,}974 \cdot \hat{x}_i = 10{,}068 - 0{,}974 \cdot 2 = 8{,}121 \text{ тис.грн•м}^2\text{/міс.}$$

На рівні значущості $\alpha = 0{,}05$ надійний інтервал прогнозу за виразом (3.32):

$$\Delta\hat{y} = 1{,}892 \cdot \sqrt{\frac{1}{7} + \frac{(\hat{x} - 5{,}04)^2}{65{,}03}} = 1{,}892 \cdot \sqrt{\frac{1}{7} + \frac{(2 - 5{,}04)^2}{65{,}03}} = 1{,}01 \text{ тис.грн•м}^2\text{/міс.}$$

Отже, для 95 зі 100 подібних об'єктів ставка оренди перебуватиме в межах $8{,}121 \pm 1{,}01$ тис. грн .•м^2/міс., або від 8020 до 8222 тис. грн за квадратний метр на місяць.

3.2.5. Складання нелінійних регресійних рівнянь

Перебіг багатьох економічних явищ та процесів відбувається за нелінійним законом. Наприклад, з мікроекономічної теорії відомий закон спадної граничної віддачі, коли збільшення факторного показника на одиницю призводить до збільшення результаційного показника на меншу кількість одиниць. За нелінійним законом зростають доходи й фінансові результати підприємства у міру збільшення обсягу виробництва продукції, виконання робіт і послуг. Також нелінійно зростає спрацьованість основних засобів: її найбільша інтенсивність припадає на перші та останні роки строку експлуатації машин й обладнання, тоді як середина терміну корисного застосування характеризується прискореним знеціненням такого активу.

Для побудови теоретичного рівняння нелінійної регресії також застосовують МНК, але спочатку теоретичну нелінійну модель шляхом алгебраїчних перетворень лівої та правої частин зводять до лінійного вигляду. Така процедура називається ***лінеаризацією***, про неї вже йшлось у підрозділі 3.2.1. Досить просто лінеаризувати гіперболічні та ступеневі моделі, для цього потрібно поділити одиницю на обидві частини регресійного рівняння або виконати логарифмування.

Зокрема, на графіку рис. 3.1,*а* наведено ступеневу регресію залежності обсягів виконаних будівельних робіт від витрат праці. Для побудови моделі потрібно насамперед лінеаризувати ступеневу модель шляхом логарифмування обох її частин.

Як відомо, логарифмування «знижує на порядок» арифметичні дії у логарифмовуваному виразі, адже логарифм – показник степеня.

- Під час множення степенів показники додають, а тому операція множення вільної константи на степінь факторної ознаки після логарифмування стане додаванням.
- Піднесення степеня до іншого степеня є еквівалентним множенню показників, а тому степеневий коефіцієнт біля факторної ознаки (***x***) стане співмножником.

Проілюструємо подані положення на прикладі степеневої регресійної моделі:

$$y = a_0 \cdot x^{a_1}. \tag{3.33}$$

Модель (3.23) лінеаризують логарифмуванням так:

$$\ln y = \ln a_0 + a_1 \cdot \ln x. \tag{3.34}$$

Звісно, замість натуральних можна виконати і логарифмування за основою, проте для обох частин рівняння мають бути застосовані однакові основи. Крім того, наприкінці розрахунків вільну константу a_0 доведеться визначати потенціюванням не за експонентою, а за основою 10.

Залежність (3.34) відрізняється від «традиційної» парної регресії ($y=a_0+a_1 \cdot x$) лише позначеннями. І саме заміна цих позначень допоможе

перетворити систему нормальних рівнянь лінійної регресії (3.13) на систему нормальних рівнянь степеневої регресії:

$$\begin{cases} \sum \ln y = \ln a_0 \cdot n + a_1 \cdot \sum \ln x \\ \sum (\ln y \cdot \ln x) = \ln a_0 \cdot \sum \ln x + a_1 \cdot \sum (\ln x)^2. \end{cases} \quad (3.35)$$

Для того щоби скласти систему 3.35, потрібно таблицю підготовчих розрахунків у вигляді таблиці, у якій спочатку обчислюють логарифми факторної та результативної ознак (відповідно ***lnx*** та ***lny***), а далі розраховують квадрати й добутки логарифмів та всі відповідні суми.

Із розв'язку системи 3.35 знаходять тільки коефіцієнт a_1 – тобто показник степеня моделі, тимчасом як, замість a_0, вільної константи-співмножника одержують лише її логарифм, $\ln a_0$. Остаточно коефіцієнт для рівняння регресії (3.33) визначають за допомогою потенціювання:

$$a_0 = exp(\ln a_0) \Rightarrow a_0 = 2{,}71828^{\ln a_0}. \quad (3.36)$$

Приклад 3.6. За даними рис. 3.1,*а* треба побудувати нелінійну степеневу регресію (3.33).

Вихідні дані та допоміжні розрахунки для системи нормальних рівнянь (3.35) зводимо в таблицю (табл. 3.6).

Таблиця 3.6

Вихідні дані та проміжні розрахунки для побудови степеневої регресійної моделі (3.33)

Пор. номер спосте-реження	Витрати праці, x_i, люд.-год	Обсяг робіт із улаштування покрівлі, y_i, м^2	ln x_i	ln y_i	(ln x_i)^2	ln x_i •ln y_i
1	5	350	1,609	5,858	2,590	9,428
2	6	480	1,792	6,174	3,210	11,062
3	7	525	1,946	6,263	3,787	12,188
4	8	520	2,079	6,254	4,324	13,004
5	8	600	2,079	6,397	4,324	13,302
6	6	390	1,792	5,966	3,210	10,690
7	5	400	1,609	5,991	2,590	9,643
8	7	490	1,946	6,194	3,787	12,054
Разом	**52**	**3755**	**14,853**	**49,098**	**27,823**	**91,371**

Система (3.35) набуде вигляду:

$$\begin{cases} \sum 49{,}098 = \ln a_0 \cdot 8 + a_1 \cdot \sum 14{,}853; \\ \sum 91{,}371 = \ln a_0 \cdot \sum 14{,}853 + a_1 \cdot \sum 27{,}823. \end{cases} \quad (3.37)$$

Для складання системи (3.37) використано такі результати розрахунків (табл. 3.6):

$$\sum \ln x = 14{,}853;\ \sum \ln y = 49{,}098;\ n = 8;\ \sum (\ln x)^2 = 27{,}823;\ \sum (\ln y \bullet \ln x) = 91{,}371.$$

Розв'язок системи 3.37: a_1**=0,87,** **ln** a_0**=4,522**. Увага: 4,522 – не коефіцієнт моделі, а тільки його логарифм. Значення вільної константи співмножника одержимо потенціюванням згідно з виразом (3.36):

$$a_0 = exp(\ln 4{,}522) \Rightarrow a_0 = 2{,}71828^{\,4{,}522} \approx 92{,}0\ м^2. \qquad (3.36)$$

Цей коефіцієнт має таку саму розмірність, як і результаційний показник. Він характеризує потенційний обсяг робіт, який можна було б виконати за 1 люд.-год.

Остаточно степенева регресійна залежність (3.33) набуде вигляду:

$$y = 92{,}0 \bullet x^{0{,}87}, \qquad (3.37)$$

де y – обсяг виконаних робіт з улаштування покрівель, м^2;

x – витрати праці робітників-будівельників, люд.год.

Отримане рівняння майже повністю збігається з наведеною на рис. 3.1,*а* формулою.

Розбіжності на рівні сотих й тисячних зумовлені нижчою точністю округлень порівняно з програмними функціями Excel. Те, як побудувати регресійне рівняння на діаграмі, викладено у підрозділі про аналіз часових рядів.

Економічне значення коефіцієнтів моделі:

- a_0***=92,0*** – вільна константа-співмножник, відображає сумарний вплив факторів інтенсифікації будівельного виробництва, відмінних від додаткових витрат праці. Цей результат можна охарактеризувати також як економічний потенціал не тільки робітників, а й усієї економічної системи підприємств, що охоплює вплив зовнішніх факторів, не включених до моделі. Так, за найкращої організації робіт, дотримання передових технологій, ощадливого використання матеріальних ресурсів, сприятливих погодних умов за 1 люд.-год можна було б улаштувати 92 м^2 покриттів;
- a_1***=0,87*** – показник степеня, відображає вплив затрат праці, тобто екстенсивного чинника, на обсяг виконаних робіт. Значення цього коефіцієнта є меншим за 1: 0,87<1, що підтверджує *закон спадної граничної продуктивності змінного фактора.* Це означає, що в разі збільшення затрат праці на 1% обсяг виконаних робіт збільшиться не на 1, а лише на 0,87% порівняно із теоретичним значенням.

Зокрема, у разі затрат праці в розмірі 5,5 люд.-год ($\hat{x}_1 = 5{,}5$) відповідно до виразу (3.37) слід сподіватися, що будуть влаштовані:

$$\hat{y}_1 = 92{,}0 \bullet \hat{x}_1^{\,0{,}87} = 92{,}0 \bullet 5{,}5^{0{,}87} = 405{,}42\ м^2\ покрівлі.$$

У разі затрат праці в розмірі 5 люд.-год ($\hat{x}_0 = 5{,}0$) за залежністю (3.37) прогнозуємо виконання робіт в обсязі

$$\hat{y}_0 = 92{,}0 \cdot \hat{x}_0^{\,0{,}87} = 92{,}0 \cdot 5{,}0^{0{,}87} = 373{,}16 \ м^2 \ покрівлі \ .$$

Таким чином, унаслідок зростання екстенсивного фактора затрат праці на 10% ($\Delta\%x = \frac{\hat{x}_1 - \hat{x}_0}{\hat{x}_0} \cdot 100\% = \frac{5{,}5 - 5{,}0}{5{,}0} \cdot 100\%$) теоретичне значення обсягу виконаних робіт зростає тільки на 8,7% ($\Delta\%y = \frac{\hat{y}_1 - \hat{y}_0}{\hat{y}_0} \cdot 100\% = \frac{405{,}42 - 373{,}16}{373{,}16} \cdot 100\%$). Іншими словами, приріст результативного показника відстає від приросту факторного у 0,87 раза $\frac{\Delta\%y}{\Delta\%x} = \frac{8{,}7\%}{10\%} = 0{,}87$.

Якби показник степеня був вищим за 1 (***a_1>1***), то внаслідок зростання затрат праці на 1% обсяг виконаних робіт зростав би більш ніж на 1% порівняно із теоретичним значенням за попереднього рівня фактора. Рівність a_1 одиниці є свідченням прямо пропорційної залежності обсягу робіт від витрат праці, що є відповідним постійній віддачі від масштабу.

3.3. ТРЕНДОВИЙ АНАЛІЗ

3.3.1. ОСНОВНІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Трендовий аналіз дає змогу виявити загальну тенденцію зміни показника в часі. Для обчислення теоретичного значення рівня у деякий момент часу t застосовуються різні функції. Розрахунки теоретичних значень та прогнозів виконують за допомогою рівнянь (аналітичних моделей). Ці рівняння дуже подібні до парної регресії, проте незалежною змінною у них є не ***x***, а ***t*** – порядковий номер періоду спостереження чи прогнозу. Відповідно у найпростішому випадку рівняння лінії тренду буде лінійним і матиме вигляд

$$y = a_0 + a_1 \cdot t. \qquad (3.38)$$

Параметри a_0 та a_1 трендового рівняння обчислюють аналогічно до параметрів нахилу та вільної константи парної лінійної регресії. Проте, зважаючи на просту послідовність значень *t*, розрахунки параметрів трендового рівняння можна суттєво спростити. Для цього використовують ***центрування***, тобто переміщують точку відліку на середину часового проміжку так, щоби сума номерів (тобто значень ***t***) і відповідно середнє значення ***t*** дорівнювали нулю.

Приклад 3.7. Виконаємо центрування послідовності спостережень за забезпеченістю населення супутниковими антенами за 2004 – 2016 роки. Кожному зі спостережень слід надати номер, причому нумерувати можна різними способами (табл.3.7).

Таблиця 3.7

Приклад центрування початку відліку часового ряду

Рік	**2004**	**2006**	**2008**	**2010**	**2012**	**2014**	**2016**	**Сума**	**Середнє**
Номер періоду за «звичайною» нумерацією, **без центрування**	0	1	2	3	4	5	6	21	3,5
Центрований номер, t	-3	-2	-1	0	1	2	3	$\sum t = 0$	$\bar{t} = 0$

- Якщо кількість рівнів ряду – непарне число, нуль, тобто точка відліку, є відповідною рівню посередині інтервалу;
- Якщо кількість рівнів ряду парна, рівні нумерують так, щоб два нульових значення, тобто подвійна точка відліку, були відповідні двом рівням посередині інтервалу.

Загалом інтервал вихідних даних, спостережуваних у попередні моменти часу, називають **ретроспективним.** Оскільки внаслідок центрування $\sum t = 0$, $\bar{t} = \frac{\sum t}{n} = 0$, отже, $\bar{t}^2 = \left(\frac{\sum t}{n}\right)^2 = 0$, формули для розрахунку оцінок коефіцієнтів трендової моделі матимуть простіший вигляд порівняно із рівняннями (3.14) та (3.15):

- **a_1** – нахил лінії тренду обчислюють за формулою

$$a_1 = \frac{\overline{y \bullet t}}{\overline{t^2}} \qquad (3.39)$$

замість формули $a_1 = \frac{\overline{y \bullet t} - \bar{y} \bullet \bar{t}}{\overline{t^2} - \bar{t}^2}$, якщо б не виконувалось центрування;

- **a_0** – вільна константа, що являє собою середню у ретроспективній вибірці спостережень**:**

$$a_0 = \bar{y} \qquad (3.40)$$

замість формули $a_0 = \bar{y} - a_1 \bullet \bar{t}$, якщо б не виконувалось центрування.

Тобто справедливим є припущення, що значення показника упродовж часу спочатку рівномірно наближається, а потім рівномірно віддаляється від деякого середнього рівня.

Змістова характеристика коефіцієнтів моделі полягає у такому:

- a_1 – нахил лінії тренду показує, на скільки одиниць в середньому змінюється значення рівня ряду в кожен наступний період часу;

- a_0 – вільна константа, відповідна середньому рівневі показника упродовж ретроспективного періоду.

Подібно до парної регресійної моделі якість трендового рівняння характеризує коефіцієнт достовірності апроксимації (коефіцієнт детермінації):

$$R^2 = 1 - \frac{\frac{\sum(y-\hat{y})^2}{n-2}}{\frac{\sum(y-\bar{y})^2}{n-1}}, \text{ або } R^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_y^2}. \quad (3.41)$$

Що більше R^2, тобто що ближче R^2 до 1, тим більш достовірно трендові рівняння відображають фактичну варіацію показника у часі. Навпаки, що менше R^2, тобто що ближче R^2 до 0, тим гірша трендова модель.

Для прогнозування за допомогою лінії тренду у рівняння слід підставити порядковий номер прогнозного періоду, тобто $\hat{t}$. При цьому беруть до уваги нумерацію періодів, за допомогою якої отримано параметри моделі.

Приклад 3.8. Для випадку центрування 2018 р. матиме порядковий номер 4 ($\hat{t} = 4$), а 2020 р. відповідно матиме номер 5 ($\hat{t} = 5$). ***Рівняння лінійного тренду*** дає змогу обчислити ***точкову оцінку прогнозу:*** $\hat{y} = a_0 + a_1 \bullet \hat{t}$. Однак номер прогнозного періоду не потрапляє до вибірки ретроспективних даних, за якими отримано трендову модель. Тому точна точкова оцінка буде некоректною, натомість краще визначати ***інтервальну*** оцінку $\hat{y} \pm \Delta y$. Надійний інтервал прогнозу розраховують аналогічно до інтервалу парної регресії, проте розрахунок дещо спрощується завдяки центруванню:

$$\Delta y = t^*_{\frac{\alpha}{2};\nu=n-2} \bullet \sigma_\varepsilon \bullet \sqrt{1 + \frac{1}{n} + \frac{\hat{t}^2}{\sum t^2}}. \quad (3.42)$$

3.3.2. Комп'ютеризація трендових розрахунків

За допомогою програми Excel можна побудувати лінію тренду та скласти рівняння лінії тренду, паралельно отримуючи показник детермінації з

мінімальними витратами часу. Для цього потрібно виконати дії в такій послідовності:

1. Ввести значення рівнів ряду в клітинки табличного процесора (рис. 3.5).
2. За діапазоном клітинок побудувати графік (рис.3.5).
3. Виділити графік й у контекстному меню обрати пункт Добавить линию тренда (чи Линия тренда) (рис. 3.6).
4. У вікні налаштувань активувати (рис. 3.7):

- «ЛИНЕЙНАЯ»,
- «ПОКАЗЫВАТЬ УРАВНЕНИЕ НА ДИАГРАММЕ»,
- «ПОМЕТИСТЬ НА ДИАГРАММЕ ВЕЛИЧИНУ ДОСТОВЕРНОСТИ АППРОКСИМАЦИИ R^2».

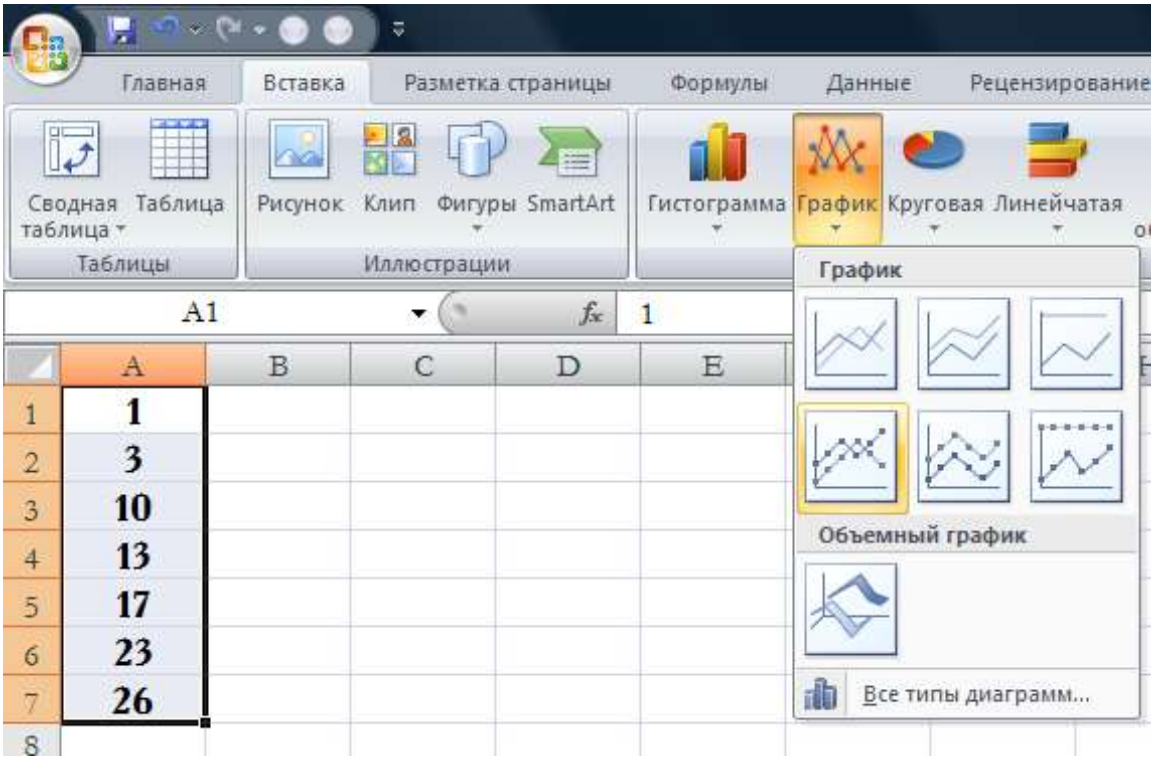


Рис. 3.5. Підготовка даних для графічного трендового аналізу (етапи 1, 2)

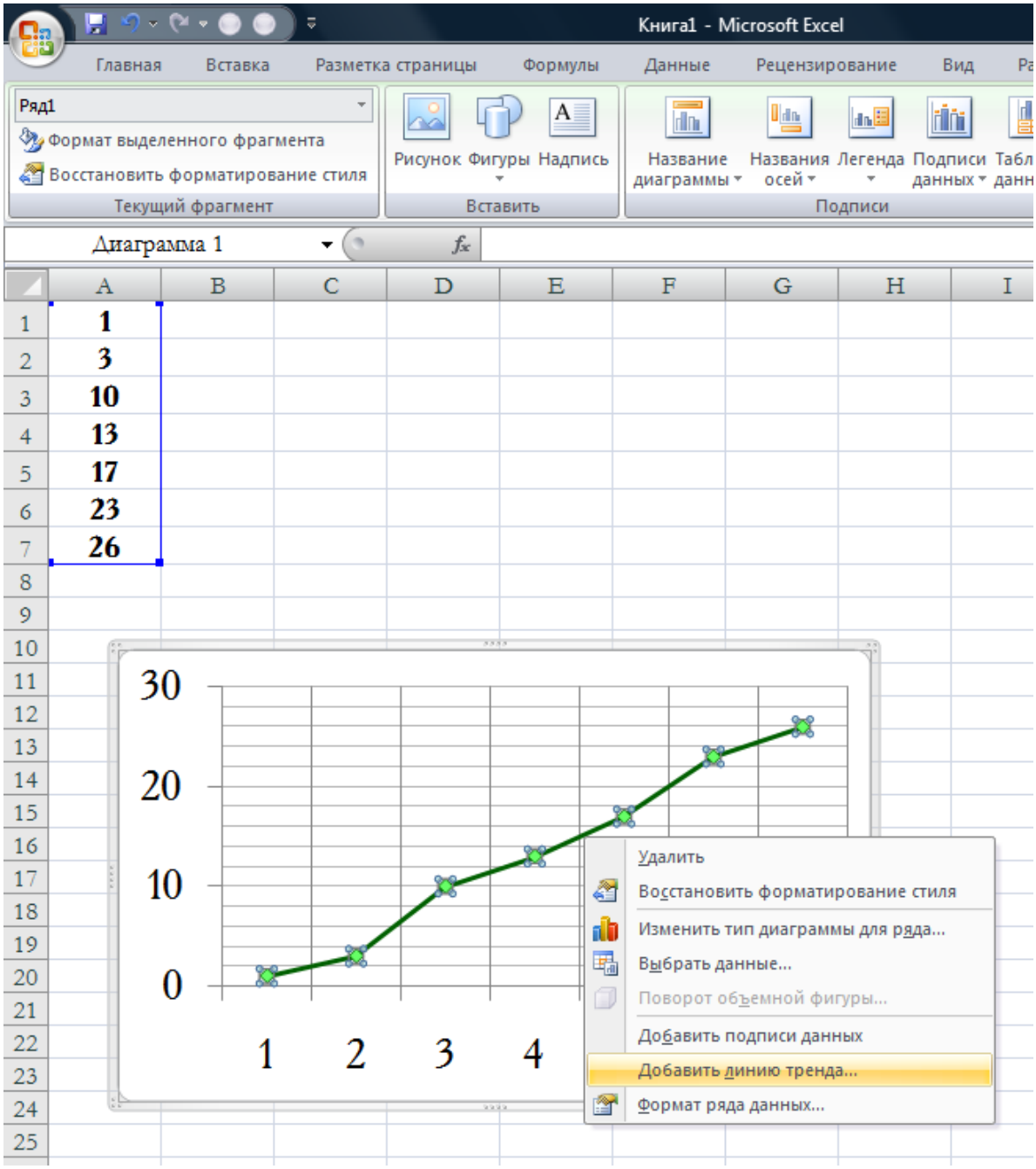


Рис. 3.6. Побудова графіка трендової лінії (етап 3)

У більш пізніх версіях Excel інтерфейс діалогових вікон несуттєво відрізняється від того, що на рис. 3.7, однак у них наявні зазначені налаштування, потрібно лише посунути нижче повзунок у смузі прокручування.

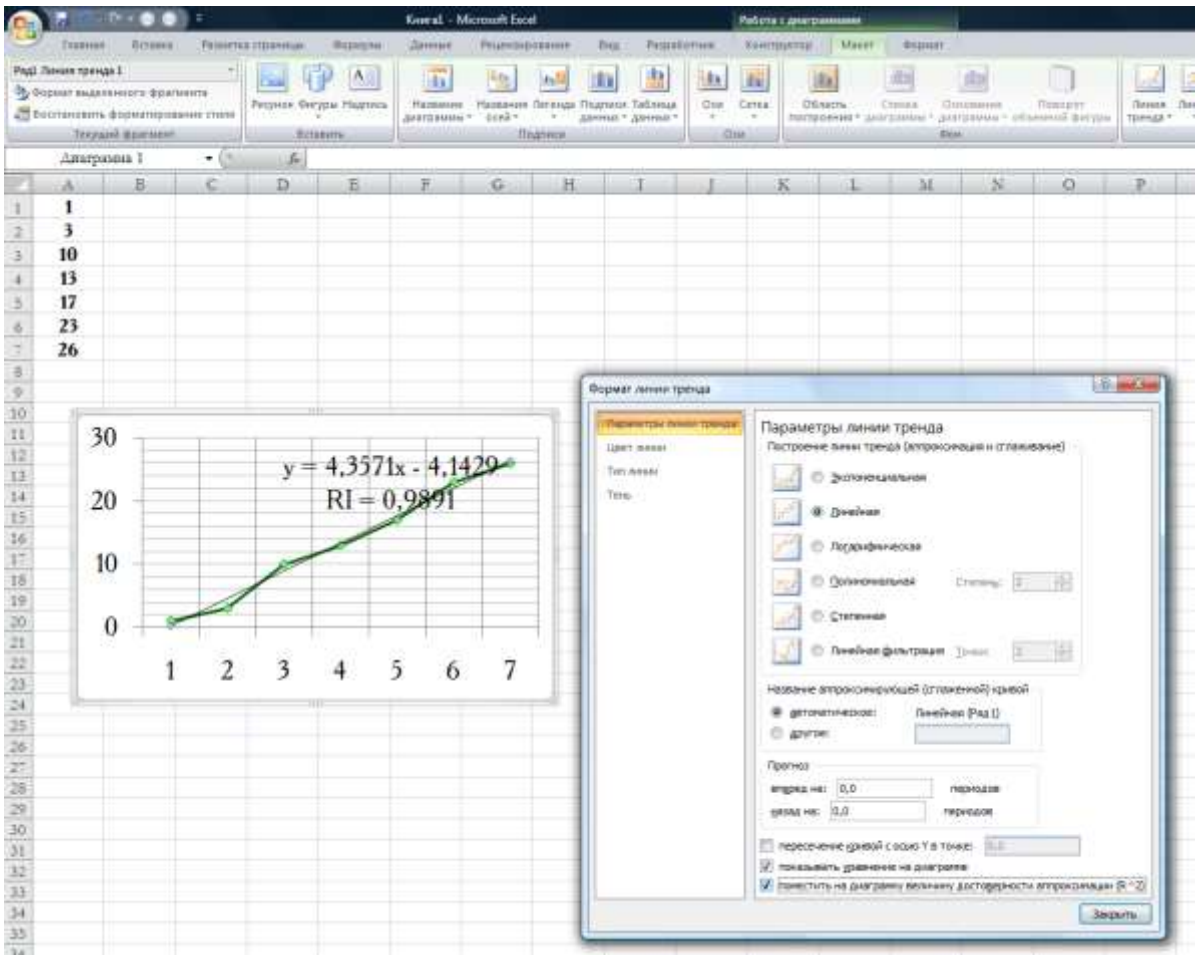


Рис. 3.7. Налаштування параметрів трендової лінії (етап 4)

У програмі порядковий номер прогнозного періоду позначений як *x*, а параметри трендової лінії розраховано без центрування. Тому значення вільної константи є не відповідним середньому рівню ретроспективних даних. Втім, оцінка нахилу трендової лінії не залежить від початку відліку.

Приклад 3.9. Зразок виконання трендового аналізу за даними прикладів 3.6, 3.7.

Для того щоби спростити розрахунки оцінок коефіцієнтів трендової моделі, виконують центрування. Вихідні дані та проміжні розрахунки зводимо до табл. 3.8.

За даними табл. 3.8 та формулами (3.39), (3.40) оцінки коефіцієнтів трендового лінійного рівняння будуть такими:

$$a_1 = \frac{\overline{y \bullet t}}{\overline{t^2}} = \frac{17{,}4}{4} \approx 4{,}36\,;$$

$$a_0 = \overline{y} = 13{,}3\,.$$

Таблиця 3.8

Вихідні дані та проміжні розрахунки
для визначення параметрів лінійного трендового рівняння (3.38)

Рік	2004	2006	2008	2010	2012	2014	2016	Сума	Середнє
Вихідні дані									
супутникові антени, y	**1**	**3**	**10**	**13**	**17**	**23**	**26**	$\sum y = 93$	$\bar{y} = 13{,}3 = a_0$
Центрований номер, t	—3	—2	—1	0	1	2	3	$\sum t = 0$	$\bar{t} = 0$
Розрахункові показники									
y•t	—3	—6	—10	0	17	46	78	$\sum y \cdot t = 122$	$\overline{y \cdot t} = 17{,}4$
t^2	9	4	1	0	1	4	9	$\sum t^2 = 28$	$\bar{t}^2 = 4$
y^2	1	9	100	169	289	529	676	$\sum y^2 = 1773$	$\bar{y}^2 = 253{,}3$
$\hat{y} = 13{,}3 + 4{,}36 \cdot \hat{t}$	**0**	**5**	**9**	**13**	**18**	**22**	**26**	$\sum \hat{y} = 93$	$\bar{y} = 13{,}3 = a_0$
$\varepsilon = (y - \hat{y})$	1	—2	1	0	—1	1	0	0	0
$\varepsilon^2 = (y - \hat{y})^2$	1	4	1	0	1	1	0	$\sum \varepsilon^2 = 8$	$\sigma_\varepsilon^2 = \frac{\sum \varepsilon^2}{n-2} = \frac{8}{7-2} = \frac{8}{5} = 1{,}6$

Таким чином, тенденція зміни забезпеченості населення супутниковими антенами може бути формалізована у вигляді лінійного рівняння (3.38)

$$\hat{y} = 13{,}3 + 4{,}36 \cdot \hat{t}\,. \qquad (3.43)$$

Економічний зміст параметрів моделі (3.43) полягає у такому:

- **a_1=4,36** – нахил лінії тренду означає, що кожні два роки кількість домогосподарств, які мають супутникову антену, зростає на 4,36 на кожні 100 домогосподарств (звичайно, краще сказати, що кількість власників супутникових антен кожні два роки збільшується на 436 у розрахунку на 10 000 домогосподарств);
- **a_0=13,3** – вільна константа, яка свідчить про те, що у середньому за 2004 – 2016 рр. чисельність власників антен становила 13,3 на кожні 100 домогосподарств (або 133 на кожні 1000 домогосподарств). Іншими словами, це вплив на забезпеченість антенами інших факторів, аніж час, причому ці фактори до трендової моделі не належать.

Згідно з рівнянням лінійного тренду (3.43) $\hat{y} = 13{,}3 + 4{,}36 \cdot \hat{t}$ обчислюють точкові оцінки ретроспективних рівнів, помилки моделі та квадрати помилок. Результати розрахунків внесено до табл. 3.8, (останні три рядки таблиці).

Як показують розрахунки, дисперсія похибки (3.17) становить:

$$\sigma_\varepsilon^2 = \frac{\sum \varepsilon^2}{n - 2} = \frac{8}{7 - 2} = \frac{8}{5} = 1{,}6\,,$$

тоді стандартне відхилення похибки буде рівним (3.19):

$$\sigma_\varepsilon = \sqrt{\sigma_\varepsilon^{\,2}} = \sqrt{\frac{\sum \varepsilon^2}{n-2}} = \sqrt{\frac{8}{7-2}} = \sqrt{\frac{8}{5}} = \sqrt{1{,}6} = 1{,}265\,.$$

Враховуючи дисперсію кількості антен у домогосподарствах

$$\sigma_y^{\,2} = \left(\overline{y^2} - \overline{y}^2\right) \cdot \frac{n}{n-1} = \left(253{,}3 - 13{,}3^2\right) \cdot \frac{7}{7-1} = 76{,}78 \cdot \frac{7}{6} = 89{,}57\,,$$

визначимо показник достовірності апроксимації трендової моделі (3.41):

$$R^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_y^2} = 1 - \frac{1{,}6}{89{,}57} = 0{,}982\,.$$

Він дещо відрізняється від аналогічного показника, розрахованого в Excel на графіку трендової лінії (див. рис. 3.7) через округлення. Згідно з величиною коефіцієнта достовірності апроксимації лінійна трендова модель пояснює мінливість кількості антен у домогосподарствах на 98,2%. Решта 1,8% варіацій залишаються непоясненими, оскільки пов'язані з впливом інших факторів, аніж час.

Відповідно до встановленої лінійної трендової моделі прогноз забезпеченості населення супутниковими антенами у 2018 році ($\hat{t} = 4$) такий:

- точкова оцінка: $\hat{y}_{2018} = 13{,}3 + 4{,}36 \cdot 4 = 30{,}74 \approx 31$. Імовірно, що у 2018 р. із кожних 100 домогосподарств 31 матиме супутникову антену. Проте коректніше вказати межі можливих прогнозних оцінок;
- інтервал можливих значень прогнозу обчислено на рівні значущості 5%, тобто табличне значення коефіцієнта Стьюдента дорівнює $t^*_{\frac{\alpha}{2};\nu=n-2} = t^*_{\frac{0{,}05}{2};\nu=7-2} = t^*_{\frac{0{,}05}{2};5} = 2{,}57$. Тоді, беручи до увазі попередні розрахунки і підсумки таблиці, матимемо:

$$\Delta y = t^*_{\frac{\alpha}{2};\nu=n-2} \cdot \sigma_\varepsilon \cdot \sqrt{1 + \frac{1}{n} + \frac{\hat{t}^2}{\sum t^2}} = 2{,}57 \cdot 1{,}265 \cdot \sqrt{1 + \frac{1}{7} + \frac{4^2}{28}} = 4{,}26 \approx 4 \text{ } \textit{антени}\,.$$

Отже, варто очікувати, що забезпеченість населення супутниковими антенами у 2018 році буде в межах 31 ± 4 антени на 100 домогосподарств, іншими словами, цей показник перебуватиме в інтервалі від 27 до 35.

3.4. НЕПАРАМЕТРИЧНІ МЕТОДИ ДОСЛІДЖЕННЯ ЗВ'ЯЗКІВ В ЕКОНОМІЦІ БУДІВНИЦТВА

3.4.1. Основні теоретичні відомості

Дослідження взаємозв'язків між економічними явищами за допомогою методів аналітичного групування та кореляційно-регресійного аналізу потребує використання таких основних характеристик популяції, як середнє значення та дисперсія. Як відомо, ці характеристики є параметрами багатьох законів розподілу значень ознак. Тому методи вивчення аналітичного групування та кореляційно-регресійного аналізу називають також параметричними. Обов'язкова умова застосування таких методів – кількісна величина результаційної (для аналітичного групування) або навіть усіх ознак (для регресійного аналізу). Та якщо потрібно встановити наявність і характер зв'язку між якісними ознаками, які можуть бути виміряні не у кількісній, а в номінальній чи порядковій шкалі, параметричні методи є непридатними, адже у такому разі неможливо визначити параметри розподілу ознак, не кажучи вже про нормальність закону розподілу. З цією метою застосовують інші методи, що мають назву непараметричних методів. ***Непараметричні методи*** – це спеціальні методи статистичних досліджень, за допомогою яких можна вимірювати зв'язок між явищами та ознаками, вираженими у номінальній та порядковій шкалах, не використовуючи при цьому кількісних значень ознак, тобто параметрів розподілу. Непараметричні методи широко застосовують у соціально-психологічних дослідженнях, коли особливості реакцій, думок, поглядів чи поведінки індивідів не можна виміряти кількісно та охарактеризувати певним законом розподілу.

Шкала – набір властивостей явища і відповідних їм значень (чисел), за допомогою яких вимірюють рівень значень ознаки. Американський психолог С. Стівенс у 40-х роках XX ст. запропонував чотири типи вимірювальних шкал: номінальні, порядкові, інтервалів та відношень. Цей підхід дав змогу поєднати кількісний та якісний підходи до аналізу ознак, адже, відповідно до пропозицій Стівенса, процес вимірювання – це процес надання об'єктам чи явищам числових характеристик, визначених відповідно до певних правил. Чотири типи шкал утворюють ієрархічну послідовність, у якій кожна наступна (вища) шкала охоплює властивості попередніх (нижчих) шкал.

***Номінальна шкала* (*шкала найменувань*)** – надання об'єктам певних найменувань або числових шифрів, наприклад, це нумерація маршрутів громадського транспорту, найменування транспортних засобів («автобус», «тролейбус»). Звісно, немає сенсу виконувати арифметичні дії над номерами автобусних маршрутів або пошук середньо-статистичного виду транспорту. Втім, пріоритет у виборі того чи іншого транспортного засобу можна з'ясувати на основі деякої «моди»: «модний» транспортний засіб буде найчастіше використовуваним громадянами і матиме найбільшу частоту сукупності пасажирських перевезень.

Порядкова* (*рангова*) *шкала дає можливість встановити зв'язок між властивостями об'єктів. Вона визначає не тільки подібність елементів, але й послідовність типу «більше, ніж», «швидше, ніж», «краще, ніж»,

«дорожче, ніж» тощо: *маршрутне таксі швидше, ніж автобус*. Кожній точці (найменуванню) шкали може бути наданий певний порядковий номер – ***ранг.*** Ранг є своєрідним рейтингом об'єкта у системній сукупності досліджуваних об'єктів. Як ранг можна розглядати і бальну оцінку знань студента чи школяра. При цьому сума балів, як і розрахунок середнього бала, з погляду статистики не є коректним, адже «двійка» та «трійка» в сумі не становитимуть відмінної оцінки («п'ятірки») з будь-якого предмета.

Інтервальна шкала дає змогу кількісно виміряти інтервал між властивостями об'єктів, тобто відображає відстань між упорядкованими точками шкали. Найбільш відомі шкали – різні температурні шкали (Цельсія, Кельвіна, Фаренгейта) та календарне літочислення. У цих шкалах, як й у багатьох інших, від'ємні значення. Шкала інтервалів дає можливість застосовувати більшість математико-статистичних методів для обробки та аналізу даних, отриманих за її допомогою: можна використовувати всі міри центральної тенденції та розсіювання, коефіцієнт кореляції Пірсона та ін. Проте шкали не можна використовувати для визначення пропорцій і відношень. Наприклад, зниження температури від 30 до 15 °С не можна трактувати як похолодання у два рази. Отже, нагромадження кількісних значень інтервалів не утворює нового об'єкта.

Шкала відношень визначає співвідношення між властивостями об'єктів, наприклад, *«поїздка на літаку в три рази дорожча за поїздку на автобусі».* Ця шкала дає змогу створити новий об'єкт шляхом об'єднання окремих об'єктів. Тому результат об'єднання, виражений у шкалі відношень, не повинен залежати ані від точки відліку, ані від одиниць виміру. У цій шкалі немає від'ємних значень, як не може бути від'ємного об'єму чи маси. Шкала відношень дає можливість виконувати всі арифметичні дії, використовувати всі параметричні методи, найбільш поширена вона у фізиці та техніці. Її обмеження для соціальних наук зумовлене тим, що не можна в результаті об'єднання характеристик окремих індивідів отримати суб'єкт із новою характеристикою, хоч би то рівень агресивності, тривожності, інтелекту чи задоволеності покупкою або сервісом.

Аналіз зв'язків між номінальними ознаками виконують на основі даних, систематизованих у вигляді спеціальних таблиць співзалежності. Числа з клітинок таких таблиць – це частоти деяких умовних та безумовних розподілів, на основі яких обчислюють спеціальні показники, за якими складають висновок про наявність чи брак зв'язків між номінальними й порядковими змінними. Розрізняють два типи таблиць співзалежності:

- **чотириклітинкові (чотирипольні) таблиці**, які складають за двома альтернативними ознаками, з яких одна може бути якісною, а друга – кількісною, або ж обидві можуть бути якісними. За кожною з ознак утворюють по дві альтернативні групи, призначені для виявлення й

оцінювання кількісного зв'язку між ознаками. Макет таблиці подано на рис. 3.8;

- **багатоклітинкові таблиці, або таблиці контингенції**, які складаються за однією альтернативною ознакою, а за другою – якісною, або кількісною, які набувають більш ніж два значення і призначені для оцінювання щільності зв'язку між ними.

Ознаки		Ознака № 1		Разом
		частоти альтернативи № 1	частоти альтернативи № 2	
Ознака № 2	частоти альтернативи № 1	a	b	a+b
	частоти альтернативи № 2	c	d	c+d
Разом		a+c	b+d	a+b+c+d=n

Рис. 3.8. Макет чотириклітинкової таблиці співзалежності

У таблицях взаємної співзалежності підсумковий рядок, який містить ряд розподілу одиниць сукупності за результативною ознакою ***y***, незалежно від факторної ознаки ***x*** називається ***безумовним розподілом***. Інші рядки таблиці взаємної співзалежності, які містять частоти розподілів за ознакою ***y*** для певних фіксованих значень ознаки ***ч*** називаються ***умовними розподілами***. Зміна умовних розподілів у певному напрямі свідчить про наявність стохастичного зв'язку між ***y*** та ***x***.

3.4.2. Дослідження зв'язку за допомогою чотириклітинкових таблиць

Для вимірювання щільності зв'язку за даними чотириклітинкових таблиць співзалежності англійським статистиком Е.Дж. Юлом (1871 – 1951) запропоновано два показники:

- коефіцієнт асоціації:

$$Q = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}; \tag{3.44}$$

- коефіцієнт колігації:

$$w = \frac{\sqrt{a \cdot d} - \sqrt{b \cdot c}}{\sqrt{a \cdot d} + \sqrt{b \cdot c}}. \tag{3.45}$$

Під час аналізу чотириклітинкових таблиць співзалежності для характеристики міри «відносного ризику» обчислюють відношення шансів W, тобто відношення перехресних добутків частот:

$$W = \frac{a \bullet d}{b \bullet c}. \qquad (3.46)$$

Якщо добутки діагональних частот однакові, то зв'язку між досліджуваними альтернативними ознаками немає.

Знак коефіцієнтів колігації та асоціації визначають за знаком чисельника, обидва коефіцієнти мають такі самі властивості, що й коефіцієнт парної кореляції. Зв'язок між досліджуваними ознаками підтверджується, якщо їх абсолютне значення є не меншим за 0,5 (|Q|>0,5; |w|>0,5). Та якщо абсолютне значення коефіцієнтів асоціації та колімації є меншим, ніж 0,5, такий зв'язок вважають неістотним.

Варто зазначити, що, по-перше, коефіцієнт колігації завжди є меншим за коефіцієнт асоціації: w<Q. По-друге, обидва зазначені коефіцієнти мають суттєвий недолік – якщо в одній із чотирьох клітин немає частоти, тобто вона дорівнює нулеві, величина коефіцієнтів асоціації та колігації за модулем завжди дорівнює одиниці. Отже, обидва коефіцієнти перебільшують оцінку щільності зв'язку. Для того щ**об** и усунути зазначений недолік, використовують коефіцієнт контингенції, який обчислюють за формулою

$$K = \frac{a \bullet d - b \bullet c}{\sqrt{(a+b) \bullet (b+d) \bullet (d+c) \bullet (c+a)}}. \qquad (3.47)$$

Значення коефіцієнта контингенції (*K*) завжди є меншим за коефіцієнти асоціації (*Q*) та колігації Юла (*w*), а тому зв'язок між досліджуваними ознаками підтверджується, якщо коефіцієнт контингенції за модулем перевищує 0,3 (|K|>0,3). Отже, коефіцієнт контингенції дає більш «стриману» оцінку тісноти зв'язку. Перевірку істотності коефіцієнта асоціації К виконують за допомогою критерію Пірсона (хі-квадрат χ^2), статистична характеристика якого функціонально пов'язана з коефіцієнтом контингенції таким чином:

$$\chi^2 = 2 \bullet n \bullet |K|. \qquad (3.48)$$

У випадку істотності коефіцієнта контингенції розрахункове значення має бути більшим за табличне, обчислене з деяким обґрунтованим дослідником рівнем значущості α та кількістю ступенів вільності $\nu = 1 = 2 - 1$. Кількість ступенів вільності є різницею між кількістю категорій ознак та одиниці: оскільки обидві ознаки є альтернативними і містять кількість категорій, рівну 2, то кількість ступенів вільності дорівнює одиниці.

Критичне значення $\left(\chi^2_{\alpha;\nu}\right)$* обчислюють за допомогою статистичних таблиць (зразок такої таблиці наведено у розділі «Методи перевірки статистичних гіпотез»), або програмними засобами, наприклад Excel шляхом введення до клітинки функції **=ХИ2ОБР(**$\alpha;\nu$**)**. У більш пізніх англомовних версіях Excel цю функцію слід задавати так: **=CHISQ.INV(**$1-\alpha;\nu$**),** тобто першим аргументом є надійність – різниця між 1 й довірчою імовірністю).

Приклад 3.10. Зразок аналізу чотириклітинкової таблиці взаємної співзалежності між успішністю студентів у вивченні вищої математики та заняттями живописом. Вихідні дані зведемо у таблицю 3.9.

Таблиця 3.9

Розподіл студентів-респондентів за цикавістю до занять живописом та успішністю у вивчення вищої математики

Розподіл студентів за ознаками, осіб	Успішність вивчення вищої математики		Разом
	Середній бал з навчальних модулів низький (меншій за 74)	Середній бал з навчальних модулів високий (74 і вище)	
навчались у художній студії або брали приватні уроки живопису	a=17	b=8	**25**
не навчались у художній студії і не брали приватних уроків живопису	c=12	d=3	**15**
Разом	**29**	**11**	**40**

Відповідно до вхідних даних (табл. 3.9), кількість опитаних респондентів – 40 осіб (n=40), решта позначень наведено у табл. 3.9.

– коефіцієнт асоціації (3.44):

$$Q = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c} = \frac{17 \cdot 3 - 8 \cdot 12}{17 \cdot 3 + 8 \cdot 12} = \frac{51 - 96}{51 + 96} = \frac{-45}{147} = -0{,}306 \Rightarrow |-0{,}306| < 0{,}5;$$

– коефіцієнт колігації (3.45):

$$w = \frac{\sqrt{a \cdot d} - \sqrt{b \cdot c}}{\sqrt{a \cdot d} + \sqrt{b \cdot c}} = \frac{\sqrt{17 \cdot 3} - \sqrt{8 \cdot 12}}{\sqrt{17 \cdot 3} + \sqrt{8 \cdot 12}} = \frac{7{,}14 - 9{,}80}{7{,}14 + 9{,}80} = -0{,}157 \Rightarrow |-0{,}157| < 0{,}5;$$

– відношення шансів (3.46):

$$W = \frac{a \cdot d}{b \cdot c} = \frac{17 \cdot 3}{8 \cdot 12} = \frac{51}{96} = 0{,}53;$$

– коефіцієнт контингенції (3.47):

$$K = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (b+d) \cdot (d+c) \cdot (c+a)}} = \frac{17 \cdot 3 - 8 \cdot 12}{\sqrt{(17+8) \cdot (8+3) \cdot (3+12) \cdot (12+17)}} = \frac{-45}{\sqrt{25 \cdot 11 \cdot 15 \cdot 29}} =$$

$$= \frac{-45}{\sqrt{119625}} = \frac{-45}{345{,}87} = -0{,}13 < 0{,}3.$$

Статистика Хі-квадрат (3.48) становить:

$$\chi^2 = 2 \cdot n \cdot |K| \cdot = 2 \cdot 40 \cdot |-0{,}13| = 10{,}41.$$

Табличне значення критерію Пірсона на рівні значущості $\alpha = 0{,}05$ та кількості ступенів вільності $\nu = 1$ становить $\left(\chi^2_{0{,}05;1}\right)^* = 3{,}841$. Цей результат отримано за допомогою Excel в результаті введення до клітинки функції =ХИ2ОБР(0,05;1) або =CHISQ.INV(0,95;1). Оскільки табличне значення критерію

Пірсона виявилось меншим за розрахункове ($10{,}41 > 3{,}841 \Rightarrow \chi^2 > \left(\chi^2_{0{,}05;1}\right)^*$), коефіцієнти контингенції є статистично значущим на рівні 0,95.

Результати розрахунків коефіцієнтів асоціації, колігації, відношення шансів та контингенції дають підстави для висновків з імовірністю 95%:

- коефіцієнти асоціації та колігації за модулем є значно меншими за 0,5, тобто зв'язку між систематичними заняттями живописом та успішністю у вивчені вищої математики немає;
- відношення шансів, що становить W=0,53, означає, що шанс натрапити на студента, який би цікавився живописом та мав труднощі з вивченням вищої математики, є майже вдвічі меншим, ніж студента, байдужого до занять живописом, але вищою за середню оцінкою з вищої математики. Іншими словами, студенти, що цікавляться живописом і не виявляють значних здібностей у математиці, трапляються в 1,88 ($\frac{1}{0{,}53}$) раза рідше, ніж студенти, які не мають проблем з вивченням вищої математики і при цьому не виявляють жодної цікавості до занять живописом;
- коефіцієнт контингенції, менший за модулем, ніж 0,3, також свідчить про брак зв'язку між систематичними заняттями живописом та успішністю у вивчені вищої математики;
- від'ємні значення коефіцієнтів асоціації, колігації та контингенції створюють передумови для припущення про протилежну спрямованість результатів занять живописом і вищою математикою: інтенсивність одного із видів діяльності негативно позначається на результатах іншої діяльності.

3.4.3. ВИВЧЕННЯ ЗВ'ЯЗКУ ЗА ДОПОМОГОЮ БАГАТОКЛІТИНКОВИХ ТАБЛИЦЬ

Якщо таблиці контингенції складено на основі комбінаційного розподілу одиниць сукупності за двома взаємопов'язаними ознаками (***x*** та ***y***), як кількісними, так і номінальними, і при цьому кожною з них визначено більше, ніж дві групи (категорії відповідно m_x та m_y), тоді для оцінювання зв'язку між ознаками ***x*** та ***y*** застосовують коефіцієнти контингенції (взаємної співзалежності) Чупрова і Крамера. Обидва коефіцієнти мають розподіл, близький до розподілу Пірсона (Хі-квадрат, χ^2), а кількість ступенів є добутком зменшеної на одиницю кількості категорій за кожною з ознак: $\nu = (m_x - 1) \cdot (m_y - 1)$. Значущість коефіцієнтів контингенції Чупрова і Крамера буде підтверджена, коли розрахункове значення χ^2 перевищить табличне: $\chi^2 > \left(\chi^2_{\alpha;\nu}\right)^*$. Розрахункове значення обчислюють так:

$$\chi^2 = n \cdot \left| \sum_i \sum_j \frac{(f_{ij})^2}{f_{\Sigma i} \cdot f_{\Sigma j}} - 1 \right|, \qquad (3.49)$$

де n – обсяг досліджуваної вибірки, як сума всіх частот (значень клітинок таблиці з частотами $n = \Sigma f_{ij}$;

f_{ij} – фактична частота *j*-го стовпчика за ***i***-м рядком;

$f_{\Sigma i}$ – підсумкова частота за ***i***-м рядком;

$f_{\Sigma j}$ – підсумкова частота за *j*-м рядком.

Отриманий показник χ^2 у разі його статистичної значущості використовують для обчислення коефіцієнтів контингенції за такими формулами:

– коефіцієнт Чупрова:

$$C_{ч} = \sqrt{\frac{\chi^2}{n \bullet \sqrt{(m_x - 1) \bullet (m_y - 1)}}} = \sqrt{\frac{n \bullet \left(\sum_i \sum_j \frac{(f_{ij})^2}{f_{\Sigma i} \bullet f_{\Sigma j}} - 1 \right)}{n \bullet \sqrt{(m_x - 1) \bullet (m_y - 1)}}} ; \quad (3.50)$$

– коефіцієнт Крамера:

$$C_{к} = \sqrt{\frac{\chi^2}{n \bullet (m_{\min} - 1)}} = \sqrt{\frac{n \bullet \left(\sum_i \sum_j \frac{(f_{ij})^2}{f_{\Sigma i} \bullet f_{\Sigma j}} - 1 \right)}{n \bullet (m_{\min} - 1)}} , \quad (3.51)$$

де m_{min} – мінімальна кількість груп за факторною (*x*) чи результативною (*y*) ознакою.

Якщо кількість категорій за обома ознаками однакова, значення коефіцієнтів Чупрова і Крамера збігаються.

Приклад 3.11. Зразок аналізу багатоклітинкової таблиці контингенції, після перевірки наявності зв’язку між інтенсивністю занять спортом та обсягом споживання молочної продукції. Вихідні дані зведемо у таблицю 3.10.

Таблиця 3.10

Розподіл респондентів за інтенсивністю занять сортом та споживання молочної продукції

Інтенсивність занять спортом	**Інтенсивність споживання молочної продукції**		**Разом**
	Низька, менша, ніж 1 кг на тиждень	**Висока, понад 1 кг на тиждень**	
Низька, менша, ніж 2 год на тиждень	8	7	**15**
Середня, понад 2 до 5 год на тиждень	5	5	**10**
Висока, понад 5 год на тиждень	2	8	**10**
Разом	**15**	**20**	**35**

Відповідно до вхідних даних кількість опитаних респондентів – 35 осіб (n=35), кількість категорій за факторною ознакою x (інтенсивність занять спортом) – 3, $\boldsymbol{m_x=3}$, а за результативною ознакою y (інтенсивність споживання молочної продукції) – 2, $\boldsymbol{m_y=2}$, тобто $m_x \neq m_y$. Мінімальна кількість груп за факторною (x) чи результативною (y) ознакою дорівнює:

$$\boldsymbol{m}_{\mathbf{min}} = \min(3;2)=2.$$

Статистика Хі-квадрат (3.49) становить:

$$\chi^2 = n \cdot \left| \sum_i \sum_j \frac{(f_{ij})^2}{f_{\Sigma i} \cdot f_{\Sigma j}} - 1 \right| = 35 \cdot \left| \frac{8^2}{15 \cdot 15} + \frac{7^2}{15 \cdot 20} + \frac{5^2}{10 \cdot 15} + \frac{5^2}{10 \cdot 20} + \frac{2^2}{10 \cdot 15} + \frac{8^2}{10 \cdot 20} - 1 \right| =$$

$$= 35 \cdot |0{,}1814 - 1| = 35 \cdot 0{,}818 = 28{,}63.$$

Табличне значення критерію Пірсона на рівні значущості 0,05 та кількості ступенів вільності $\nu = (m_x - 1) \cdot (m_y - 1) = (3-1) \cdot (2-1) = 2$ дорівнює $\left(\chi^2_{0{,}05;2}\right)^* = 5{,}991$. Цей результат отримано в Excel в результаті введення до клітинки функції =ХИ2ОБР(0,05;2), або функції =CHISQ.INV(0.95;2). Оскільки табличне значення критерію Пірсона виявилось меншим за розрахункове ($28{,}63 > 5{,}991 \Rightarrow \chi^2 > \left(\chi^2_{0{,}05;2}\right)^*$), варто зробити висновок про наявність зв'язку між інтенсивністю занять сортом та споживанням молочної продукції. Це дає підстави для розрахунку решти коефіцієнтів.

Коефіцієнти контингенції Чупрова (3.50) та Крамера (3.51) з імовірністю 0,95 дорівнюють:

– коефіцієнт Чупрова (3.50)**:**

$$C_ч = \sqrt{\frac{\chi^2}{n \cdot \sqrt{(m_x - 1) \cdot (m_y - 1)}}} = \sqrt{\frac{28{,}63}{35 \cdot \sqrt{(3-1) \cdot (2-1)}}} = 0{,}761.$$

Це означає, що у 95 випадках зі 100 є стійкий зв'язок між інтенсивністю занять спортом та споживанням молочної продукції;

– коефіцієнт Крамера (3.51):

$$C_к = \sqrt{\frac{\chi^2}{n \cdot (m_{\min} - 1)}} = \sqrt{\frac{28{,}63}{35 \cdot (2-1)}} = 0{,}904.$$

Отриманий результат означає, що у 95 випадках зі 100 є майже пропорційний зв'язок між інтенсивністю занять спортом та споживанням молочної продукції.

Через те, що таблиця вхідних даних не є квадратною (кількість категорій за x та y не збігається), значення коефіцієнтів Чупрова і Крамера також не збігаються. Оскільки у випадку неквадратних таблиць знаменник у коефіцієнта Крамера завжди менший за знаменник коефіцієнта Чупрова, коефіцієнт Крамера завищує щільність зв'язку. Таким чином, остаточним є висновок про наявність суттєвого зв'язку між інтенсивністю занять спортом і

споживанням молочної продукції з імовірністю 95%, тоді як результат про майже пропорційний зв’язок слід відхилити як завищений.

ЗАВДАННЯ ДЛЯ САМОСТІЙНОГО ОПРАЦЮВАННЯ МАТЕРІАЛУ

Увага! У визначенні числових показників для розрахунково-аналітичних завдань:

- **h*** – **остання цифра** номера паспорта;
- **g*** – **передостання цифра** номера паспорта.

Якщо біля числового показника немає буквених доданків чи співмножників, такий показник є спільним для всіх варіантів.

Надійність рішення верифікації в усіх розрахунках обрати згідно з варіантом й табл. 3.11.

Таблиця 3.11

Рівень статистичної значущості розрахунків

Варіант, h*	**2;5;8;0**	**3;6;9**	**1;4;7**
Імовірність помилки, α	**0,1**	**0,05**	**0,025**

1. За даними всіх спостережень задачі 3 розділу 2 (табл. 2.15) визначіть:

– коефіцієнт парної кореляції між стажем роботи та продуктивністю праці робітників. Перевірити значущість отриманого коефіцієнта;
– лінійну регресійну залежність продуктивності праці від стажу роботи робітників. Дати економічну інтерпретацію результатів. Оцінити якість моделі на основі коефіцієнта детермінації та F-критерію;
– встановити інтервальні оцінки парної лінійної регресії;
– знайти межі надійної зони регресії. Встановити математичне очікування продуктивності праці (змінного виробітку) робітника зі стажем роботи вісім років та межі довірчих інтервалів його виробітку;
– скласти рівняння залежності продуктивності праці від досвіду роботи робітників у вигляді степеневої регресії.

2. Складіть стохастичну модель для розрахунку потреби в адміністративно-управлінському персоналі, беручи до уваги структуру персоналу сімох найкращих будівельних підприємств, що працюють на регіональному ринку. Вихідні дані подано в табл. 3.12. Дайте економічну інтерпретацію моделі та встановіть статистичну значущість отриманої залежності.

Таблиця 3.12

Дані спостережень структури персоналу будівельних підприємств

Чисельність адміністративно-управлінського персоналу, осіб, y	11	9	7	4+h	15	10—h	6
Кількість робітників, осіб, x	39+3•h*	52	48	23+5•g*	50	50—2•h*	15

3. Обчисліть середньомісячний індекс цін на блага за 10 місяців 20-го року та за період згідно з вашим варіантом. Вихідні дані наведено у табл. 3.13. Скористайтеся формулою середнього геометричного. Вид блага також обирайте відповідно до варіанта. Визначте параметри лінійного та степеневого тренду, Наведіть розрахунки у вигляді системи нормальних рівнянь. Розрахуйте показник достовірності апроксимації та сформулюйте висновок щодо точності кожного з типів тренду. Яка формула більш точно описує тенденцію зміни цін у 20-му році? Дайте графічну інтерпретацію.

Таблиця 3.13

Динаміка цін на товари і послуги у 20XX-х роках

Варіант, h*	Благо	Січень	Лютий	Березень	Квітень	Травень	Червень	Липень	Серпень	Вересень	Жовтень
g* – парне (0,2,4,6,8), період лютий – серпень (включно)											
0	Хліб	102,9	106,2	122,7	106,2	100,6	100,3	100,0	99,7	99,9	100,1
1	Макаронні вироби	102,4	105,8	120,4	110,7	101,3	99,7	99,3	99,5	99,4	99,4
2	М’ясо та м’ясопродукти	101,4	100,2	106,2	104,4	102,1	101,2	102,0	102,9	101,1	99,3
3	Риба та продукти з риби	107,1	110,0	120,7	104,4	99,6	98,8	99,3	99,6	99,6	98,3
4	Молоко, сир та яйця	101,5	101,3	104,2	101,2	102,4	100,9	98,9	101,1	103,1	103,7
5	Молоко	101,7	101,6	105,5	102,1	99,1	99,1	99,9	101,0	102,0	103,5
6	Сир і м’який сир (творог)	101,4	102,5	104,4	103,5	101,6	100,8	100,4	100,4	101,0	101,5
7	Яйця	101,6	98,9	101,2	94,7	109,7	103,6	94,5	102,7	109,0	108,3
8	Олія та жири	102,5	104,8	116,2	104,3	101,2	100,2	100,3	100,8	101,0	100,8
9	Масло	101,1	101,5	105,4	104,7	101,9	100,3	100,1	99,9	100,6	102,3
g* – непарне (1,3,5,7,9), період травень – жовтень (включно)											
0	Олія соняшникова	105,3	110,1	132,1	104,9	101,2	100,1	99,6	100,0	99,0	97,6
1	Інші їстівні тваринні жири	99,8	100,2	104,5	101,5	99,7	99,9	101,6	103,6	105,7	105,1
2	Фрукти	113,5	117,5	131,9	102,3	109,5	97,4	93,9	82,8	101,0	101,0
3	Овочі	117,8	115,2	117,1	103,4	112,0	97,7	78,4	83,8	106,2	107,1
4	Цукор	101,9	109,5	139,2	92,8	93,3	97,9	101,9	101,4	101,5	108,2
5	Безалкогольні напої	104,5	106,6	116,9	109,6	105,4	101,8	100,6	101,0	100,9	100,3
6	Одяг і взуття	98,1	102,2	113,5	104,6	101,0	98,2	95,3	98,8	118,3	103,6
7	Меблі та предмети обстановки, килими та інші види покриттів для підлоги	103,9	108,6	110,3	101,7	100,1	99,8	100,3	100,2	100,8	100,5
8	Домашній текстиль	103,6	107,5	110,4	105,0	102,0	101,0	100,4	100,3	101,5	101,2
9	Побутова техніка	104,1	108,9	113,3	103,2	100,9	99,1	99,4	100,5	100,8	101,2

4. За даними вибірки людей у віці 30 – 60 років, наведеними у табл. 3.14**,** з'ясуйте, чи є залежність між імунітетом інженерів-проєктувальників до застудних хвороб та систематичними заняттями танцями.

Таблиця 3.14

Результати соціологічного опитування інженерів-проектувальників

Розподіл людей за ознаками, осіб	**Частота захворювань, разів на рік**	
	до 3-х включно	**понад 3**
Регулярно відвідують танцювальні студії	190+5•m*	80+10•m*
Епізодично танцюють на вечірках або не танцюють зовсім	200–10•n*	120+5•n*

Для цього обчисліть критерій χ^2**,** а також відношення шансів.

Встановіть значущість результатів та дайте їх розгорнуту інтерпретацію.

5. Встановіть на основі критерію χ^2 коефіцієнтів Чупрова і Крамера, чи є залежність між рівнем задоволеності роботою та улюбленим кольором за такими даними вибірки інженерів-кошторисників у віці 30 – 60 років (табл. 3.15)**:**

Таблиця 3.15

Результати соціологічного опитування інженерів-кошторисників

Улюблений колір	**Рівень задоволення роботою**		
	низький	**середній**	**вищий за середній**
Зелений	5	80	70+5•n*
Помаранчевий	90+5•m*	30	30+10•m*
Бузковий	200–10•n*	3	25

Встановіть значущість результатів та дайте їхню розгорнуту інтерпретацію.

ЛІТЕРАТУРА

1. Єлейко В.І. Економетричний аналіз діяльності підприємств : навч. посіб. / В.І. Єлейко, Р.Д. Бондар, М.Я. Демчишин. – Тернопіль: Навчальна книга – Богдан, 2011. – 368 с.
2. Економетричний інструментарій управління фінансовою безпекою підприємств будівництва: [моногр.] / Л.В. Сорокіна, А.Ф. Гойко, С.П. Стеценко, К.В. Ізмайлова, І.О. Шапошнікова та ін.; за наук. ред. Л.В. Сорокіної, А.Ф. Гойко. – К.: Київський національний університет будівництва і архітектури, 2017. – 404 с.
3. Круш П.В. Економіка (розрахунки фінансово-інвестиційних операцій в Excel [текст] : навч. пос. / П.В. Круш, О.В Клименко. – 3-тє вид., перероб. та допов. – К. : ЦНЛ, 2014. – 256 с.
4. Макаренко Т.І. Моделювання та прогнозування у маркетингу : навчальний посібник / Т.І. Макаренко. – К. : ЦНЛ, 2005. – 160 с.
5. Общая теория статистики : учебник / А.Я. Боярский, Л. Л. Викторова, А. М. Гольдберг и др.; под ред. А.М. Гольдберга, В.С. Козлова. – М.: Финансы и статистика, 1985. – 367 с.
6. Личковський Е.І. Вища математика. Теорія наукових досліджень у фармації та медицині : підручник / Е. І. Личковський, П. Л. Свердан. – К. : Знання, 2012. – 476 с.
7. Моторин Р.М. Статистика для економістів : навч. посіб. / Р.М. Моторин, Е.В. Чекотовський. – 3-тє вид., випр. і допов. – К. : Знання 2013. – 381 с.
8. Павлов К.В, Павлова О.М. Формування та регулювання конкурентних відносин на регіональних ринках житла України : монографія / К.В. Павлов, О.М. Павлова; Східноєвропейський національний університет імені Лесі Українки. – Луцьк : Терен, 2019. – 542 с.
9. Сорокіна Л.В. Моделі і технології управління ринковою вартістю будівельних підприємств : [Текст] / Л.В. Сорокіна. – К. : Лазурит-поліграф, 2011. – 541 с.
10. Сорокіна Л. В. Аналітична характеристика розвитку ринку житла м. Києва / Л.В. Сорокіна, А.Ф. Гойко // Шляхи підвищення ефективності будівництва в умовах формування ринкових відносин. – 2016. – Вип. 34. – С. 83-97. – Режим доступу: http://nbuv.gov.ua/UJRN/shpebfrv_2016_34_12
11. Статистика: підручник / А.В. Головач, А.М. Єріна, О.В. Козирєв та ін.: за ред. А.В. Головача, А.М. Єріної, О.В. Козирєва. – К.: Вища школа., 1993. – 623 с.
12. Справочник по прикладной статистике: в 2-х т. / под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. – М.: Финансы и статистика, 1989, 1990.
13. Статистический словарь / гл. ред. М.А. Королёв. – М.: Финансы и статистика, 1989. – 623 с.

14. Черняк О.І. Інтелектуальний аналіз даних : підручник / О.І. Черняк, П.В. Захарченко. – К.: Знання, 2014. – 599 с
15. Стеценко С.П. Ієрархічна модель оцінювання інфраструктурних ризиків підприємницької діяльності у будівництві [Електронний ресурс] / С.П. Стеценко, Т.А. Ільїна // Наукові праці НДФІ. – 2019. – Вип. 1. – С. 74-84. – Режим доступу: http://nbuv.gov.ua/UJRN/Npndfi_2019_1_7
16. Шапошнікова І.О. Аналіз часових рядів первинного ринку житлової нерухомості м. Києва / І.О. Шапошнікова // Економічний вісник університету. ДВНЗ Переяслав–Хмельницький державний педагогічний університет імені Григорія Сковороди. – 2018. – №36/1. – С.139-147 (DOI: 10.5281/zenodo.1219766).
17. Шапошнікова І.О. Сучасні тенденції розвитку первинного ринку житла [Електронний ресурс] / І.О. Шапошнікова // Вчені записки Таврійського національного університету імені В.І. Вернадського. – 2019. – Т. 30(69), № 6(1). – С. 40-47. – Серія «Економіка і управління». – Режим доступу: http://nbuv.gov.ua/UJRN/UZTNU_econ_2019_30(69)_6(1)__10
18. Щетініна, О.К. Стохастичне моделювання економічних процесів [Текст] / О.К. Щетініна, К.О. Палагута // Шляхи активізації інноваційної діяльності в освіті, науці, економіці : матеріали Всеукр. наук.-практ. конф. [м. Вінниця, 12 квітня 2016 р.] : у 2-х т. / орг. ком. : А.І. Крисоватий, З.–М. В. Задорожний, Б.В. Погріщук [та ін.]. – Вінниця : ВННІЕ ТНЕУ, 2016. – Т. 1. – С. 107–109.
19. Statistica 6. Статистический анализ данных.: [учебник] / А.А. Халафян. – 3-е изд. – М. : ООО «Бином-Пресс», 2008. – 512 с.

ІНФОРМАЦІЙНІ РЕСУРСИ

1. www.rada.gov.ua – Сервер Верховної Ради України.
2. www.minregion.gov.ua – сайт Мінрегіону України.
3. http.ukrstat.gov.ua – сайт Державної служби статистики України.
4. www.inproect.kiev.ua – сайт НПФ «Інпроект».
5. http://tk-311.1gb.ua/ – сайт технічного комітету стандартизації ТК 311 «Ціноутворення та кошторисне нормування у будівництві»
6. https://dom.ria.com/uk/ – Сайт оголошень про купівлю-продаж нерухомості
7. https://knoema.ru/atlas/topics/%d0%a2%d1%80%d0%b0%d0%bd%d1%81%d0%bf%d0%be%d1%80%d1%82/Motor-Vehicle-Sales/Car-sales – Світовий атлас статистичних даних

ДЛЯ НОТАТОК

ДЛЯ НОТАТОК

Навчальне видання

СОРОКІНА Леся Вікторівна,
ГОЙКО Анатолій Францович,
СТЕЦЕНКО Сергій Павлович та ін.

СТАТИСТИКА В УПРАВЛІННІ ЕКОНОМІКОЮ БУДІВНИЦТВА І НЕРУХОМОСТІ

Навчальний посібник

Редагування та коректура *Г.В. Кобриної*
Комп'ютерне верстання *Т.І. Кукарєвої*

Підписано до друку 22.09.2022. Формат 60 × 84 1/168
Ум. друк. арк. 9,76 . Обл.-вид. арк. 10,5.
Тираж 25 прим. Вид. № 2/І-22. Зам. № 14/1-22.

Видавець і виготовлювач
Київський національний університет будівництва і архітектури

Повітрофлотський проспект, 31, Київ, Україна, 03037

Свідоцтво про внесення до Державного реєстру
суб'єктів видавничої справи ДК № 808 від 13.02.2002 р.